DOCUMENT RESUME

ED 339 726 TM 017 603

AUTHOR Hymes, Donald L.; And Others
TITLE The Changing Face of Testing and Assessment: Problems and Solutions. AASA Critical Issues Report.
INSTITUTION American Association of School Administrators, Arlington, Va.
REPORT NO ISBN-0-87652-164-2
PUB DATE 91
NOTE 106p.
AVAILABLE FROM American Association of School Administrators, 1801 North Moore St., Arlington, VA 22209-9988 ($14.95 for 1 copy, 10% discount for 2-9 copies, 20% discount for 10 or more copies. Add $3.50 to orders of $20.00 and under for postage and handling. Orders for less than $20.00 must be accompanied by payment in full).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Academic Achievement; Accountability; Educational Assessment; Educational Change; Educational Improvement; Educational Policy; Educational Quality; *Educational Testing; Elementary Secondary Education; National Surveys; School Districts; *Standardized Tests; Student Evaluation; *Testing Problems; Test Use; Trend Analysis
IDENTIFIERS National Assessment of Educational Progress; *Performance Based Evaluation

ABSTRACT

Educational testing is at a decisive moment, as the issue of standardized testing has become one of the most hotly debated education issues of the decade. The accuracy and relevance of standardized tests have been questioned by many researchers, evidenced in the 20-year report of the National Assessment of Educational Progress. There is general agreement that improving student achievement is a top priority for educational assessment. The search for new ways to determine student achievement and to allow educational accountability has resulted in the advocacy of new approaches, such as performance assessments, that will reflect achievement more authentically. The state of testing is reviewed, and issues and concerns about test use are discussed. The search for alternatives and the ways these alternative assessments are being used are described. School districts must examine their current policies to ensure that changes result in better assessment practices. Because there is a lot more to assessment than just scores, several states have broadened the definition of school quality in response to demands for accountability. The report includes responses to a survey about testing from 239 out of 1,200 school leaders throughout the country. Twenty-four features throughout the text highlight issues of special interest. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED339726

AASA Critical Issues Report

The Changing Face of Testing and Assessment

Problems and Solutions

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

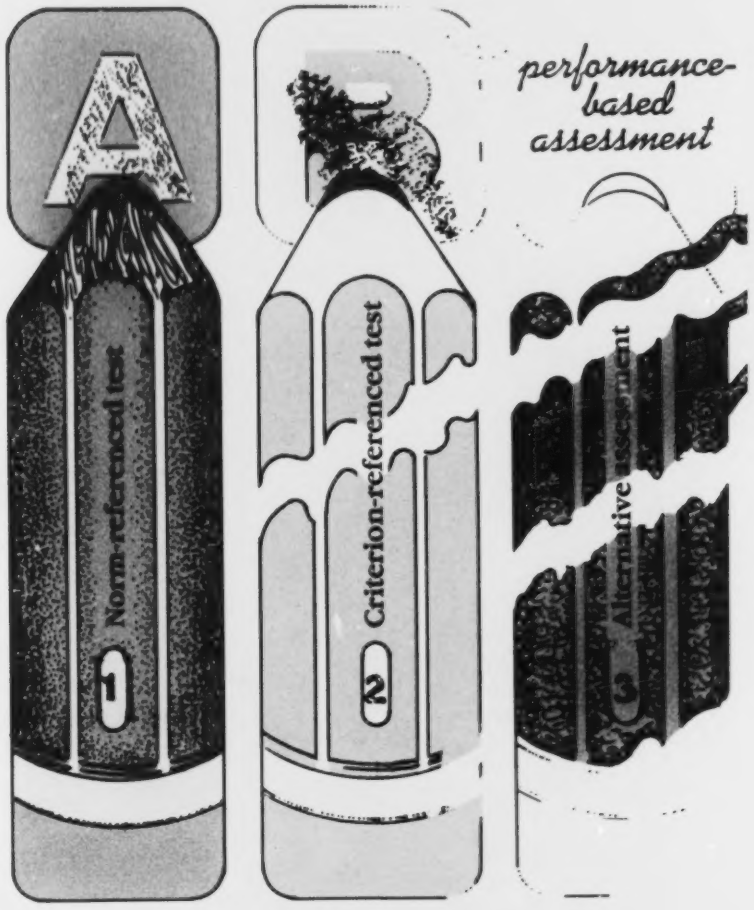
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Henry

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



By Donald L. Hymes
with Ann E. Chafin and Peggy Gonder

TM017603

 BEST COPY AVAILABLE

The AASA Critical Issues Series

Perhaps no aspect of schooling is as pervasive—and misunderstood—as testing. What was once an occasional exercise to monitor classroom instruction has taken on increasing significance in schools around the nation, where children fill in the "bubbles" more than 100 million times each year on norm-referenced standardized tests.

As the 1990s began, educators were seriously analyzing the methods by which they test student achievement. The familiar norm-referenced tests were denounced, and movements featuring performance-based assessments and portfolios burst upon the scene. A proposed national test to raise student achievement stimulated lively debate; there were cries to eliminate the reliance on "test scores" altogether and instead adopt "standards." And through it all, there was the challenge of finding tests that could enable teachers to teach better and still satisfy the public demand for accountability.

Where does the local district fit into this scenario? Where is this evolution likely to take America's schools? This Critical Issues Report attempts to find the answers by reviewing the problems and the proposed solutions. It does not take the traditional statistical approach to educational testing, but looks at it as a dynamic institution in the midst of major change.

The report was written by Donald L. Hymes (who also edits the Critical Issues series) with Ann E. Chafin, testing and evaluation specialist with the Charles County, Maryland, Public Schools and former chief of the program assessment branch of the Maryland State Department of Education; and Peggy Gonder, an education writer and public relations consultant living in Denver, Colorado. Special thanks to Chris Eichle, testing specialist for Charles County, for his help; to Bruce Hunter, AASA associate executive director for governmental relations, for his review and comments; and especially to Luann Fulbright, AASA director of communications, for her editorial guidance.

Our thanks also go to the school administrators who responded to the AASA survey conducted for this report and to the assessment specialists in the state departments of education who contributed much of the material, with special appreciation to Barbara Weiss of the California Assessment Program.

RICHARD D. MILLER
Executive Director
AASA

GARY MARX
Associate Executive Director
AASA

Copyright © 1991, American Association of School Administrators

Library of Congress Card Number: 91-61874

ISBN: 0-87652-164-2

AASA Stock No.: 021-00338

Design and Graphics: Anita F. Winfield

This report was reviewed by AASA and approved for publication.
Approval does not signify that the contents reflect the views or policies of AASA.

Price for a single copy, $14.95. Two to nine copies, 10% discount; 10 or more copies, 20% discount. Please add $3.50 to orders of $20 and under for postage and handling. Orders for less than $20 must be accompanied by payment in full. Order from AASA, 1801 North Moore St., Arlington, Va. 22209-9988; (703) 528-0700.

AASA is an equal opportunity employer.

AASA Critical Issues Report

The Changing Face of

Testing and Assessment

Problems and Solutions

By Donald L. Hymes
with Ann E. Chafin and Peggy Gonder

Developed under the editorial direction of
Donald L. Hymes
Editor, Critical Issues Reports

Contents

Preface ***Testing and Assessment:***
Evolution or Revolution? **4**

Chapter 1 ***Overview:***
The State of Testing **9**

Testing has been a mainstay in education for well over 100 years. And while there have been dramatic new developments in the field during the past two decades, the old standbys are still very much in evidence.

Chapter 2 **Issues and Concerns** **18**

It takes a lot of conviction to challenge an institution as firmly entrenched and widely used as standardized tests, but that hasn't stopped a growing band of critics from blaming them for many of education's ills.

Chapter 3 **The Search for Alternatives** **25**

Under pressure from many directions to clean up their asessment act, many members of the educational community have set out to explore new and different ways of measuring the success of the schools and their students.

Chapter 4 **Alternatives in Action** **33**

The search for more authentic assessments has seen a variety of promising programs introduced that seem to answer many of the concerns about traditional standardized testing.

Chapter 5 **Beware the Bandwagon** **49**

While the burgeoning field of authentic assessment claims many converts among its ranks, there remain many skeptics who warn against getting carried away. The answer, they suggest, may lie in a marriage between the two camps.

Chapter 6 **Assessment Begins at Home** **54**

Before buying into any of the the new techniques, or settling for the old ones, school districts should ask themselves whether they have a comprehensive assessment program, or just give tests.

Chapter 7 **The Front Line of Assessment** **63**

A working knowledge of assessment techniques is critical if teachers are to use it effectively to help students, yet training in the field is woefully inadequate.

Chapter 8 **The Politics of Assessment** 77

Charges that the achievement of the nation's students has not improved in 20 years have provoked a controversy at the national level involving the threat of national standards and a national test.

Chapter 9 **The Power of Communication** 84

To some, the word "communication" might only signify releasing test scores. But if used to its fullest extent, it can bring understanding and support from staff and community alike.

Chapter 10 **More Than Just Test Scores** 98

One of the outcomes of the reevaluation of educational assessment is the realization that there is a lot more to school performance than test scores. Responding to demands for "accountability," several states have broadened the definition of school quality.

FEATURES

National Trends in Average Proficiency 8
AASA Survey Results 15
A Concise Glossary of Assessment Terms 16
What If . . . A Vision of Educational Assessment 26
Maintaining Balance in Pittsburgh 31
Performance Assessment: Living Up to Expectations 32
Open-Ended Math: Sample Problems 35
Assessing the Essential Skills 38
Sample Scoring Rubrics in Writing and Math 40
A Sample Performance Assessment: The Soda Task 47
An Interview with Greg Anrig 53
Code of Fair Testing Practices in Education 59
Criteria for Evaluation of Assessment Systems 61
Standards for Teacher Competence 69
Sample Assessment Planning Charts 73
Student Self-Assessment Checklist 75
Guidelines for Test Item Writing 76
Two Decades of the Nation's Report Card 80
The National "Achievement" Goal 82
Is Another "National Test" on the Horizon? 83
Recruiting Parents as Partners 87
Reporting on Performance Assessment 93
Preparing the Parents 97
Maryland's New "Report Card" 100

Testing and Assessment:

Evolution or Revolution?

"Not everything that counts can be counted, and not everything that can be counted counts."

—Albert Einstein

Testing is at a crossroad.

Those who have long toiled in the obscurity of a specialty few people understood suddenly find themselves in the spotlight.

Politicians at all levels turn to testing to evaluate the health of the nation's schools. Educators at all levels, stung by recent criticism, are willing to be held accountable, but not by tests that do not have a direct bearing on student learning.

The media, which used to ignore such things as test results because they were "dull," now find ways to make them exciting.

Assessment has become a high-profile business, and those who make a living in it are feeling the heat.

"This is a critical moment in the history of assessment," said Richard J. Stiggins of the Northwest Regional Education Laboratory. "The education community must educate its members in the use and misuse of assessment. Educators must demand assessments that meet real instructional needs first and give accountability second."

Issue of the decade

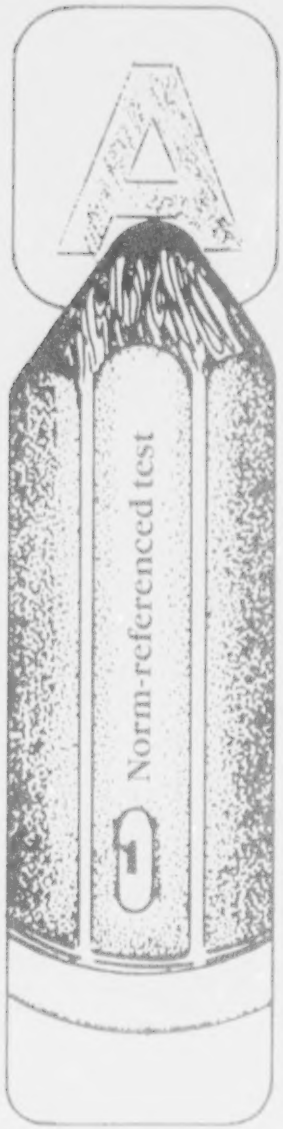
Standardized testing has become one of the most hotly debated education issues of the past decade, dating perhaps from 1980 when one of Ralph Nader's organizations took on one of the major players in the assessment game, the Educational Testing Service (ETS).

It didn't take long for other reformers to take aim at a number of sacred cows, and soon in many quarters the long-honored term "educational testing" was dismissed as an oxymoron.

But for every shrill critic there were 100 happy users. How else can we account for recent reports that well over 100 million standardized achievement and ability tests are administered each year to America's public school children? And if they are allegedly worthless, how come the growth rate in the volume of standardized testing in elementary and secondary schools has been estimated at between 10 and 20 percent *each year* over the last 40 years?

Obviously standardized tests fill a need, and they do it at a price most financially strapped school districts can afford. With pressure emanating everywhere from the White House to the corner supermarket to improve school performance, educators still turn to standardized tests to prove the quality of their reform efforts.

But how closely do standardized tests match what students are being taught—or even more important, what they should be taught? And if a goal of American education is to inspire higher order thinking skills in children, is the standard multiple-choice format the right way to assess them?



A turning point

The decade of the 90s dawned to the echo of questions like these. The hallowed institution of standardized testing was coming under increasing fire, especially as more and more classroom time was devoted to tests. Virtually everybody gave norm-referenced tests; some districts added criterion-referenced tests, while most states required minimum competency tests. Then there were IQ tests and aptitude tests. And there was the National Assessment of Educational Progress (NAEP). These congressionally mandated tests, which have been sampling student achievement for more than 20 years, became the focus of a firestorm of controversy revolving around proposals for national standards and a national test.

Why all this activity to add yet another layer on top of already burdensome testing programs? To many people at the national level, the answer was obvious: The current system is not working.

Late in 1990, NAEP issued a report summarizing the findings during its 20-year lifetime. The report, which at one point called itself "a bleak portrait of the current status of student achievement in the United States," noted that students score far below reasonable levels of competency in reading, writing, math, arithmetic, science, history and government. As for the trend: "Overall achievement levels are little different entering the 1990s than they were two decades earlier."

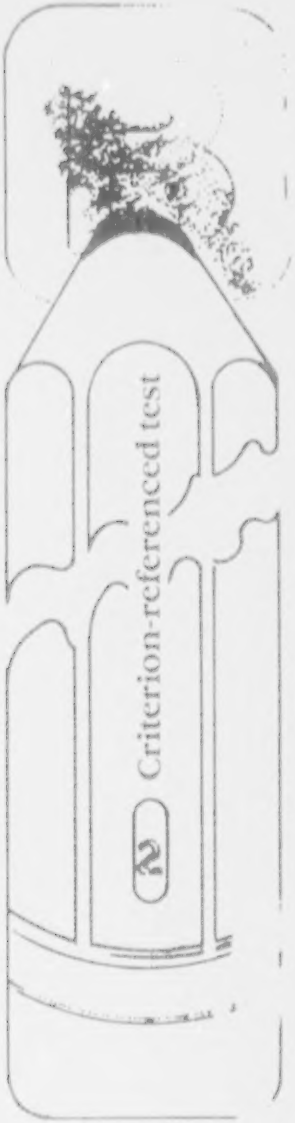
Furthermore, the report deplored the gaps in performance between white and minority students and between males and females in some subjects. It pointed out that the majority of students were not taking challenging courses in math and science, and concluded:

"Across the past 20 years, little seems to have changed in how students are taught. Despite much research suggesting better alternatives, classrooms still appear to be dominated by textbooks, teacher lectures, and short-answer activity sheets."

Changes suggested

This NAEP report, titled, "America's Challenge: Accelerating Academic Achievement," offered only general recommendations, such as these:

- More instructional time should be given to the core academic subjects.
- More "student-centered approaches" should be attempted to increase the effectiveness of instructional time.
- Parents and their children should understand the importance of spending more time on homework and learning outside of school.



- The entire nation, including parents, students, educators, policymakers, and voters, should join in a concerted effort to achieve our national goal of improved academic achievement for all students.

How will we know?

Even though many educators questioned the accuracy of the NAEP 20-year summary, everybody agreed that the top priority was improving student achievement. But that brought them back to the original question: How will we know that we are succeeding? If the answer is not traditional standardized testing programs, where else can schools turn as reform efforts cry out for more and more "data?"

Hope was on the horizon in the form of a growing movement seeking to develop alternative or "authentic" methods of assessment. That movement promised to be a high priority in the 1990s as a variety of educational reform movements gathered speed.

This hatched a chicken-and-egg issue that added another dimension: While tests are obviously needed to measure the success of a reform movement, they also can be very influential in bringing about reform. So as the conventional multiple-choice tests were scorned as barriers to reform, new forms of assessment were promoted as levers of change because they have the power to influence teachers' behavior.

As Lauren and Daniel Resnick, two authorities in the field, put it, tests become "instruments of school reform" because assessments designed to measure thinking and reasoning skills are likely to encourage schools to concentrate on teaching those same skills. Thus, they added, the current "debates are primarily about curriculum and learning goals, not about techniques of assessment."

The move to alternatives

The word "assessment" has been in the educational idiom for many years but only as a synonym for "testing." As norm-referenced tests began to lose favor in the late 1980s, "assessment" was redefined to include methods other than "testing." Before long, it often was preceded by such adjectives as "authentic" and "alternative."

Alternative assessments, whether using performance-based tests, projects, or portfolios, are fast gaining popularity among educators. By the early 1990s, more than half the states were using some type of performance assessment to evaluate student writing, while in other subjects, some form of alternative method was mandated in at least eight states and about 100 school districts.

Pilot authentic assessment programs were popping almost everywhere up to supplement, and sometimes replace, traditional norm-referenced tests. But despite the growing popularity of the movement, traditionalists cautioned against jumping on the bandwagon before seeing proof that these programs can be conducted economically on a large scale, and thus become a valid method to measure school performance.

Start in the classroom

And let us not overlook the obvious. For too long, some assessment experts cried, schools have neglected the type of assessment that has been going on from time immemorial: teachers observing their students, giving tests on what they had taught, and making decisions on the outcome.

Classroom assessment, it was stressed, is geared directly to support instruction and to provide immediate and detailed feedback to help the student and the teacher. Strong doubts were raised whether teacher-training institutions prepare their graduates to make the most of that potential. But even beyond that, it is something else again for such measurements to serve as a reliable source of schoolwide, systemwide, statewide, or even nationwide data about student performance. Before such a possibility can be considered, teachers must be trained in the art and science of assessment. This is no small task.

Classroom assessment can be of enormous value to the teacher in targeting each child's instruction. But because the strength of teacher assessment lies in its immediacy and informality, it is not a suitable source of data for district accountability purposes —at least not yet.

Two faces of testing

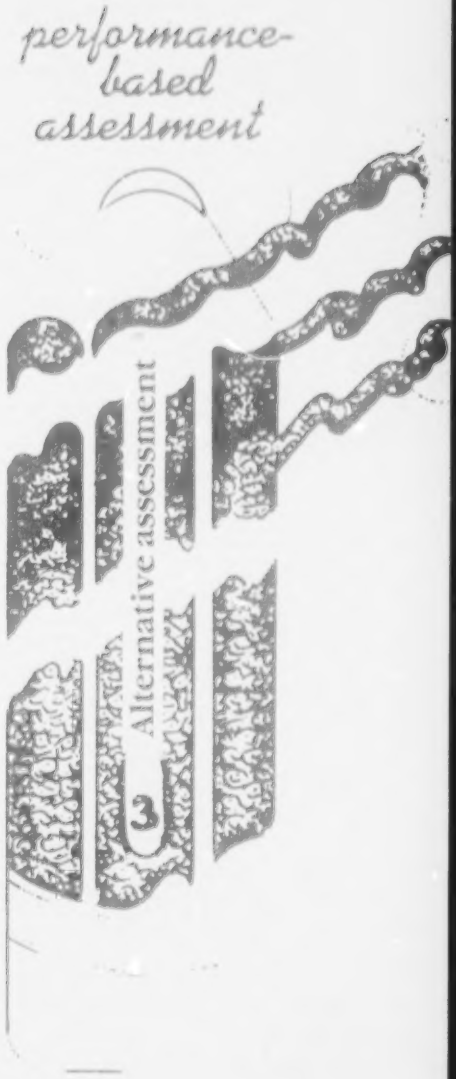
What emerges is the realization that assessment has two faces:

- Schools can test to measure how a student is learning and find ways to improve his or her instruction.
- Schools can test to monitor or "account" for student achievement from year to year, school to school, and district to district.

But can they find a way to do one without sacrificing the other?

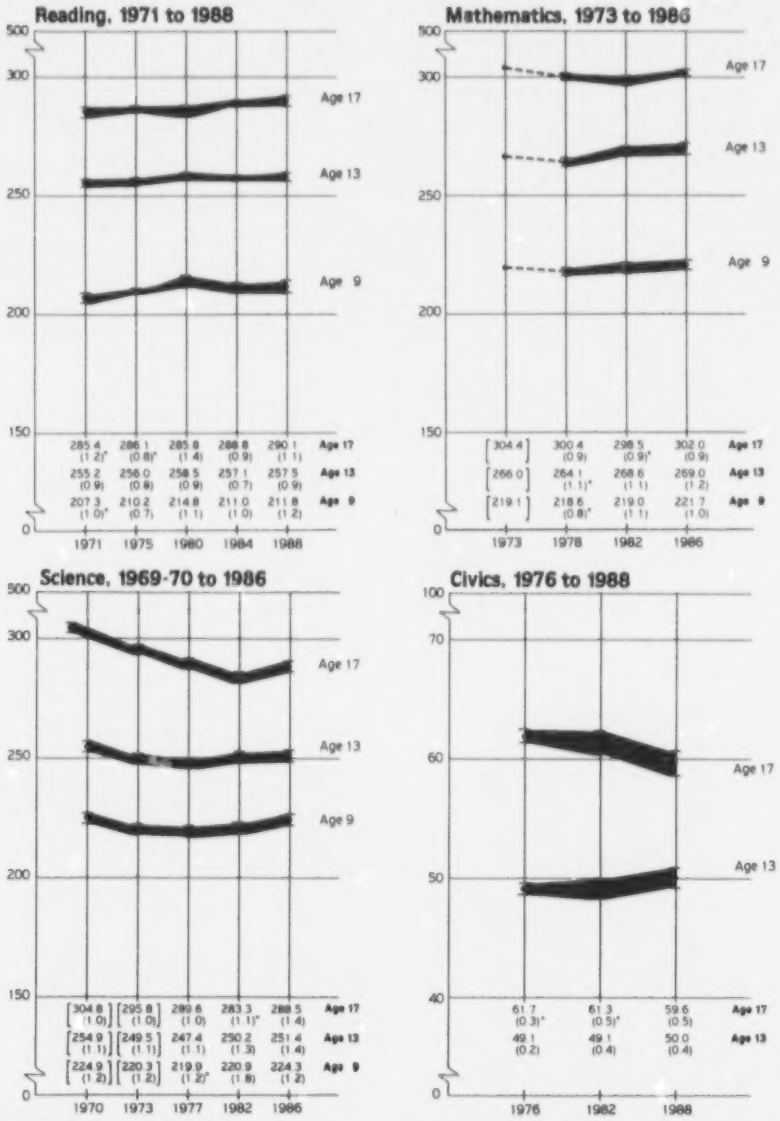
Perhaps that is why the myriad accountability movements that came into their own as the decade began were multi-faceted. They were all based on the realization that not one of the traditional measures taken by itself could paint a true picture of a student, a school, or a district. That true picture can only come from looking at the totality of a student's and school's performance—charting results on a range of assessment measures in the context of the characteristics of the school population and the community it serves.

Thus a standardized test and a performance assessment each has value, as do the rates of attendance and of graduation, and much more. True assessment will embrace all of these factors, and only then will a school and a school district be able to respond fully to the demands of "accountability."



When NAEP referred to "a bleak portrait of student achievement" in its 20-year history, it illustrated that statement with this kind of data. From "America's Challenge: Accelerating Academic Achievement," September 1990.

FIGURE **2.1** | **National Trends in Average Proficiency in Various Subject Areas: Ages 9, 13, and 17**



Standard errors are presented in parentheses. It can be said with 95 percent confidence that the average proficiency of the population of interest is within ± 2 standard errors of the estimated value.

[- - -] Extrapolations based on previous NAEP analyses.

*Statistically significant difference from the most recent assessment at the .05 level.

95% confidence interval

CHAPTER ONE

Overview:

The State of Testing

Testing has been a mainstay in education for well over 100 years. And while there have been dramatic new developments in the field during the past two decades, the old standbys are still very much in evidence.

There have been tests probably as long as there have been schools.

According to the dictionary, a test is "a series of questions, problems, etc., intended to measure the extent of knowledge, aptitudes, intelligence, and other mental traits." In that same dictionary, the definition of the word "assessment" deals only with taxes. It entered the vernacular of education only about two decades ago because "testing" had begun to take on too narrow a connotation, at a time when new and different methods were being developed to evaluate students and programs.

Perhaps soon the dictionary will define "assessment" this way: "A variety of methods and techniques used by educators to measure student knowledge, skills, and other traits. One of these techniques is 'testing.'" Until then, while it may offend some purists, the words will frequently be used interchangeably.

While there are many purposes for testing, there is basically only one reason to test: To gather the clearest, most precise, consistent, and meaningful data possible to answer questions about student performance. In the classroom, it helps the teacher teach better. As part of a districtwide effort, uses of testing information can range from defending budget increases to deciding where to buy a house, and — yes — to designing instructional programs to help children learn better.

Ideally, assessment should never be imposed from above, apart from the curriculum, but instead should be an integral part of instruction. The information belongs first in the teachers' hands so they can decide on the steps to take to continue each child's education. When it is available to principals, patterns of strengths and weaknesses in the school's instructional program can be investigated. Superintendents and school boards can use it to decide if major curricular changes are necessary.

While it may not answer all the questions, a solid job of gathering test and assessment data tells educators the current status of student performance so that teaching plans, instructional programs, and curricular changes will address real needs. If the information the decision makers get is unreliable, unclear, biased, or irrelevant, then the decisions they make will be that way too. And if the tests themselves are unreliable, unclear, biased, or irrelevant, the system is doomed from the start.

With that caveat foremost in the minds of education leaders, schools are being compelled to play a more significant role than ever before in preparing young people for the future. How well are they equipped to succeed? And how will we know if they do?

In the beginning there was testing . . .

Way back in 1854, the story goes, the Boston school trustees were concerned about the quality of instruction their students were receiving. They hit upon the idea of administering a written test to supplement the traditional oral examinations. According to historians, that may have been the first recorded incidence of a standardized test in the United States.

Whether this test actually proved anything about the quality of instruction in Boston is debatable. In fact, according to another story, 20 years later in nearby Quincy, the school board examined its students in person and found, to its chagrin, that they lacked many of the skills of reading and writing for which they had received high marks on their tests. The board hired a Col. Francis Parker to resolve the problem. The first thing he did was to prohibit teachers from specifically helping their students to pass the test. He concentrated instead on the process of education — what children and teachers did in the classroom — something a test failed to measure.

Nevertheless, the early part of the 20th Century saw the spread of standardized tests for a variety of purposes. The label "standardized" means they are designed for a single

examination to be given to a large group of students under similar conditions. Then, as now, they featured a multiple-choice format for easy administration and grading. Some, called "IQ" tests for "intelligence quotient," were used to sort out students who needed special programs or to assign children into "appropriate" tracks as early as elementary school. Appropriate or not, many remained in those tracks for the rest of their lives.

The Army used an adaptation of the IQ test on all recruits during World War I, and based on those test scores, concluded that the performance of high school graduates was distressingly low. This led to self-examination by the nation's school leaders, who hired the test developers to diagnose the problems and suggest the solutions. They did it all with standardized tests.

The concept of testing "achievement" was born in the 1920s, using the same multiple-choice technique. These tests were, on the whole, given sparingly to America's school children to help administrators monitor certain programs. A tremendous growth in the use of these tests, however, began in the 1960s because of two factors: (1) The development of computers made scoring fast and cheap; and (2) school districts turned to them as a quick and dirty way to report the success of federal compensatory education programs.

Tests tied to reform

That growth continued at a rapid pace through the 1970s as state legislatures enacted "accountability" legislation to accompany increased school funding. And it continued unabated through the 1980s, especially in response to national reports critical of the state of American education, such as *A Nation at Risk: The Imperative for Educational Reform* in 1983.

The 80s was a pivotal decade in the field of educational assessment, according to Edward D. Roeber, supervisor of the Michigan Educational Assessment Program, who is co-director of the Association of State Assessment Programs. For the first time, he said, the general public and local and state legislators became directly involved in discussions about the quality of schools, and the need for accurate data on student achievement became a priority. Many states launched large-scale assessment programs for the first time, and as the 90s began, the number of states without mandated testing programs was declining rapidly.

In addition, the purposes of these new assessment programs grew to encompass new programs and activities other than judging the effectiveness of schools. For example, Roeber said, they became factors in merit pay plans or school accreditation. And they probed more deeply than before into the traditional areas of reading and mathematics, and expanded to other subject areas as well, like science, social studies, and writing.

"Nearly every large reform effort of the last few years has either mandated a new form of testing or expanded use of existing testing," Chris Pipho of the Education Commission of the States observed in 1988. The question of whether that testing has had any real impact on the quality of education was the central issue in the debate that marked the beginning of the 1990s.

The norm-referenced achievement test

By far the most popular of the standardized tests are the norm-referenced achievement tests (NRTs). They are administered to children in virtually every school district in the nation more than 100 million times a year at a cost nationwide of between $700 and $900 million. The nationwide survey of school administrators conducted for this report showed that 95 percent of the respondents administer such tests; 94 percent consider them important in policy making and curriculum development; and 86 percent use them as part of their district's accountability program.

While a few very large school districts have developed their own achievement tests, the vast majority of those being used are one of the following: California Achievement Test (CAT); Iowa Test of Basic Skills (ITBS); Stanford Achievement Test (SAT); Comprehensive Test of Basic Skills (CTBS); and Metropolitan Achievement Test (MAT or Met). All are purchased from commercial publishers, for whom standardized tests are a major revenue item.

There are two key concepts embedded in these tests, which form the basis of the controversy: They are "achievement" tests, meaning that they attempt to measure how much a student has learned in a particular subject; and they are "norm-referenced," meaning that the local scores are compared against those of a so-called "norm group" —a nationwide sampling of students who have already been given the test by the publishers. These norms can be as much as ten years old.

The advantages of norm-referenced tests

Despite the growing controversy, the users of these tests, and their developers, maintain they have many advantages. Paul L. Williams, director of research and measurement for CTB/McGraw-Hill, one of the major test publishers in the

nation, summarized it this way: "Among the strengths of norm-referenced tests is the ability to measure a broad range of outcomes in a cost-effective, time-effective manner. This is in addition to their considerable strengths of reliability and validity."

The great majority of school districts are successfully using NRTs to take a periodic look at their students' achievement levels in certain basic skills. They watch for data on trends to signal the need for changes in their instructional programs.

Depending on what a school is trying to find out, Williams said, a norm-referenced test can be the best instrument it can use:

- If the top 5 percent of a class must be identified for advanced class enrollment, an NRT is ideal.
- If a school district has a fine remedial reading program but can afford to offer it to only 10 percent of the fifth grade, a norm-referenced test of reading skills will select those who need it most. Because of the way the tests are constructed, the students are spread out on the scale so that differences are easier to see and the selection is facilitated.
- If the objective is to predict success, an NRT is best. Distinguishing among the knowledge, ability, and skill level of students is best done with a measure that compares student scores rather than one that tells what a student knows or doesn't know.
- And if you want to ask the question, "Has instruction made a difference?" Williams added, you will find that alternative methods are far more cumbersome and still cannot tap the range of content a norm-referenced test can.

The bottom line

But probably the main advantage of the NRT was illustrated by this question asked by one administrator on the AASA survey; it was typical of many others: "If we don't use standardized test scores, what will we use? Does a small school district have the resources to develop an alternative to a national norm-referenced test?"

Cost effectiveness. In these days of shrinking budgets, the cost-effectiveness of nationally standardized tests is a major boon to most local school districts. They can, in effect, get accountability for pennies a pupil. The alternatives are far more expensive. Even while defending their programs as well worth the investment, such pioneers in authentic assessment measures as Dale Carlson, director of the California Assessment Program, will concede that the cost differential can be as much as five times per pupil.

Objectivity. These tests are considered objective in that scoring is generally done by a machine and is not based upon subjective judgments of individuals grading the papers.

Familiarity. There is something to be said for using a vehicle that has been in place for decades. The staff is familiar with the process and the results are reported as numbers that are easy to understand and to track. As an accountability instrument, the norm-referenced tests are simple and to the point, and if the same test is used year after year, a trend is easy to discern. Once in a great while there may be trouble explaining a sharp change in scores because of a periodic renorming, but unless the numbers plunge way below average, most districts can ride it out.

Criterion-referenced tests

Because of growing concern over the proliferation of norm-referenced testing, criterion-referenced tests (CRTs) are gradually supplementing and even supplanting NRTs in many areas as a means of gauging student performance. Rather than compare one student with another, a CRT measures whether a child has mastered specific learning objectives.

The test a teacher writes for use in the classroom is criterion-referenced, but it is a one-of-a-kind. If a test is to measure overall school or district performance, it must be standardized—given in the same form under the same conditions to a large number of persons to elicit comparable data. Such tests are being developed by states, by districts themselves, and by test publishers for districts that do not want to spend the time or the money to produce their own.

Understanding the difference

The following analogy may help lay people understand the difference in concept between a norm-referenced and criterion-referenced test:

Virtually every high school fields a football team regardless of the potential talent that shows up for tryouts each year. Rarely if ever do we hear the announcement, "Northside High School will not have a football team this year because no one is good enough." The process is to select the best of the group of athletes who try out for the team. This could mean a very painful autumn for the school and a lot of grumbling from the boosters, but there is still a team.

That is a classic example of a norm-referenced measure. Of the students enrolled in the school, these are the best football players. They may not be able to pass, run or tackle, but they are the best of the group.

Imagine if the same kind of selection procedure were used to license airplane pilots. Then the announcement might sound like this: "None of our graduates is able to

land very well, but these are the best of the class so they get their wings. Have a nice flight."

The passengers don't care one bit if their pilot was the valedictorian of the flight school. They do care, however, that the captain has sufficiently mastered the skills of takeoff, navigating, and landing to guarantee a safe trip.

What is called for here is a criterion-referenced measure that allows a specified set of skills to be demonstrated, not a measure that chooses the best of the bunch.

The major advantage

One of the major selling points of the criterion-referenced approach is that the information gained from these tests can be used in direct instruction to help the student attain a core of basic skills. In other words, if the essential knowledge were detailed, and how much and what of it our students knew were measured, then the school could teach them what was left.

When standardized CRTs are used in place of NRTs to determine student achievement on a school-by-school or district-by-district basis, it requires that a uniform curriculum be strictly enforced across the board. This is the case, for example, with such statewide efforts as the Missouri Mastery Program, in which the state established strict curricular goals for the schools and developed the tests to assess whether the student achieved them.

"Among the strengths of norm-referenced tests is the ability to measure a broad range of outcomes in a cost-effective, time-effective manner."

—Paul L. Williams, CTB/McGraw Hill

Not only multiple choice

Many critics of standardized testing object to CRTs because in many cases they are multiple-choice, which, they contend, leads to many abuses. But for the sake of definition, it should be clear that unlike NRTs, criterion-referenced tests need not be multiple-choice. As an example, there are writing assessments that specify the characteristics of writing that are to be demonstrated. Those are criterion-referenced. When students are given a project with directions that detail the types of evidence that must be gathered to support a hypothesis, the resulting evaluation is criterion-referenced.

Criterion-referenced tests assess a specified body of knowledge, skills or understanding to determine what has been learned and what else needs to be taught. Such movements as Mastery Learning and Outcome Based Education rely on criterion-referenced measures to decide when a student is ready to move on. Virtually all the new alternative assessment techniques being developed around the nation are, by definition, criterion-referenced. It is unusual, however, to see such tests referred to as CRTs. Generally, that term is reserved for the more traditional paper-and-pencil instruments.

Intelligence tests

This granddaddy of all standardized tests has been around since before the turn of the century, and in that period it has been both glorified and vilified. It is not, however, dead. Such tests as the Wechsler Intelligence Scales for Children, and the Stanford-Binet Intelligence Scale are reportedly given by 58 percent of the districts surveyed by AASA. Many of these respondents, however, noted that they are used only after a child is referred to a psychologist for a diagnosis of school problems.

Individually administered IQ tests like these have more varied tasks and provide more opportunities for professionals to observe behaviors that may indicate the causes of problems. But some schools still conduct regular, group-administered multiple-choice intelligence testing to help select students for giften programs, and to frame expectations for student achievement, both when they begin school and when they enter high school.

The early days of "compensatory education" in the 1960s used these tests to determine which children needed extra help. In recent years, however, this practice has diminished markedly in the face of criticism that this type of test can erroneously predict low achievement and doom some children to bottom tracks and fewer opportunities. Opponents argue that these tests are based upon the assumption that intelligence is a single, measurable trait rather than a variety of abilities. Their use also assumes that "intelligence" is fixed and permanent. Many people disagree, citing studies that show many things, like coaching, the test environment, and even a good breakfast, can affect IQ scores.

Intelligence testing will probably continue, but it has taken a back seat to achievement testing—especially the kind that measures thinking skills.

Minimum competency tests

"Minimum competency" is a generic term for a variety of standardized testing programs designed to determine whether students have attained a specified level of proficiency, and thus are worthy of promotion or graduation.

Tests of his type, which can carry such diverse labels as "functional" and "proficiency," became popular in the mid-1970s and continue to be widely used today. In fact, in about two-thirds of the states, such tests are used to track student progress. In most cases, they are written and administered under the auspices of the state education department. In the AASA survey, 51 percent of the respondents reported that they administer minimum competency tests, and 34.5 percent of all the districts surveyed said they use the results to make decisions on promotion and graduation. Interestingly, at the same time, 53 percent of those surveyed disapprove of this practice, which mirrors the checkered past of the minimum competency testing movement.

Basic skills

In the late 1970s and early 1980s, many states and districts rushed to develop and administer tests of basic skills to assure parents and the community that their children were at least "minimally" qualified to go on to the next grade or to receive a high school diploma. In many cases these movements were prompted by outcries from the community alleging that high school graduates were only marginally literate and otherwise unprepared for work. For this reason, the responsibility of specifying the "basics" to be covered in the tests was often turned over to the community through a survey asking for input on what competencies students should have.

At times, the objectives were as concrete as "Students must know how old one must be to serve as President of the United States." At other times, they were as abstract as "Students will be able to compute in real-life settings."

The predominant format of these tests was multiple-choice, which caused many of the instruments—whether at the state, district, or school level—to appear to measure rote, trivial knowledge. Critics used this perception to demand changes in assessment, as reflected elsewhere in this report. Some of the newer tests, especially in writing assessment, were performance-based, but that didn't entirely quiet the critics.

Setting the standard

The problem many assessment experts have is not with minimum competency tests as much as with the standards. While they were developed to represent the "floor," they often wind up being the "ceiling." And that, they say, means locking students in an inch-high room.

By definition, a standard is arbitrary. Somebody, whether an individual or a group, decides what level of performance is to be "passing," or "good," or "excellent" based upon judgment, experience, and a gut feeling. When one eastern state introduced its minimum competency test in writing a decade ago, the assessment staff recommended setting the passing standard at 70 percent. When one of the participants in the meeting remarked that a neighboring state's standard was 70 percent (on an entirely different test, of course), passing was changed on the spot to 80 percent.

Signal needs early

Among the first of the statewide minimum competency movements was the series of "Functional" tests in Maryland, given since the mid-1970s to all ninth graders in reading, and later expanded to writing, mathematics, and citizenship. Students get several opportunities to pass, but they cannot get a high school diploma until they do. While these tests have been criticized by some districts in the state as too "easy," they served to flag students who needed help early enough in their high school careers to permit remediation.

Among the most recent additions to the genre was the series of ninth grade "Proficiency" tests launched by the state of Ohio in late 1990. All ninth graders, except those exempted by a handicap, must pass these tests, also in reading, writing, mathematics and citizenship, to get a high school diploma. As in Maryland, all but the writing tests are multiple choice. In Ohio, however, students are required to perform multiple operations as well as demonstrate complex thinking in response to many of the questions. Plans called for similar tests to be given to 12th graders starting in 1993. A variety of high school certificates and diplomas will be awarded depending on which of the tests students pass.

Pragmatic approach

Whatever their label, many educators still criticize minimum competency testing for emphasizing "passing" rather than earning the highest possible score. But others take a more pragmatic approach. In response to the survey question, "Do you approve of the idea of using minimum competency tests as a basis for promotion or graduation?" Don Davenport, assistant superintendent in Henryetta, Oklahoma, answered yes and added: "Schools, administrators or teachers should not be allowed to be sued for students' failure. This threat is being waved in our face now." N. Kendall Pottorff, superintendent of School District 201 in Aledo, Illinois, also answered yes and was a little more succinct: "If developed wisely and with a good attorney."

It is likely that the 1990s will see more and more of the traditional minimum competency tests supplanted by more authentic assessments. Still, most educational assessment specialists concede that they learned a great deal from this movement over the past 20 years. Primarily, it has forced schools to look at their curricula and the delivery of instruction based on the results.

"If we don't use standardized test scores, what will we use? Does a small school district have the resources to develop an alternative to a national norm-referenced test?"

—Anonymous.

National Assessment of Educational Progress

This is a Congressionally mandated project created 20 years ago to measure the scholastic achievement of students in elementary, middle, and high school. Since 1969-70, tests have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields—resulting in more than 200 reports. Currently, NAEP is sampling students' knowledge and skills in math, reading, writing, science, and civics at the ages of 9, 13, and 17. While the reports traditionally have published national averages, beginning in 1990 NAEP's math assessment will enable state-by-state comparisons.

Up to this point, local school districts have been able to make little practical use of these tests and in some cases hardly acknowledge them. The primary reason, critics say, is that they sample too few students who are not especially motivated to perform well on them, and the results are released far too late to do anything with the data.

But as a motivator for federal government officials to decry a lack of improvement by the nation's youngsters and threaten further involvement in public education, NAEP became the focus of a burgeoning controversy in 1990. As a result, proposals were introduced for a national test and national standards, which will be covered in Chapter 8.

Aptitude Tests

Aptitude tests are given by 39 percent of the school districts surveyed for this report. By definition, tests of student aptitude are administered to predict performance according to a specific set of criteria before making a placement or beginning instruction. These tests, which in some ways resemble intelligence tests in their makeup and their use, are commonly used to make decisions on advanced placement or gifted and talented assignments.

In addition to that 39 percent, the most widely used tests of this type administered in virtually every community are those designed to predict high school students' success in college. These are the Scholastic Aptitude Test (SAT) and the American College Test (ACT), taken by more than 2 million 11th and 12th graders.

Colleges set standards for performance on these tests for those young men and women who want to enter their institutions. Because the results are so important, elaborate preparation courses are offered by schools and commercial agencies, and bookstores are full of handbooks and guides.

What is important in considering such aptitude tests as the SAT and the ACT is that they measure only those factors for which they were created—in this case the student's potential ability to succeed in college. They do not pretend to cover everything taught in a comprehensive high school program and thus cannot be used to evaluate it. Yet misinterpretations of this type have led to controversy in many communities and on a national level. In addition, the practice of comparing average SAT or ACT scores from one school or community to another can be highly misleading because the test is voluntary. The percentage of students who take the test has a direct impact on the average score, and these figures vary widely from school to school and community to community.

Survey Results

As with all AASA Critical Issues Reports school leaders throughout the country are surveyed to determine their practices and their feelings. For this report on testing, 1,200 questionnaires were distributed, and 239 were returned. In a separate survey, less than half of the state departments of education returned their questionnaires.

While the responses are not to be considered statistically significant, they serve as an indicator, and provide a pool of examples of past, current or intended practices in school systems.

Of the respondents, 6 percent were urban, 33 percent suburban and 60 percent rural.

Types of tests administered

- Norm-referenced tests: 95 percent. (The most frequently mentioned were the Iowa Test of Basic Skills, the CAT and CTBS by CTB/McGraw-Hill, and the Stanford Achievement Series.)
- Criterion-referenced tests: 51.5 percent.
- Aptitude tests: 39 percent.

Changes in testing practices

More than half of the respondents—53 percent—have made significant changes in their programs over the past five years, most commonly the addition of a criterion-referenced component.

When asked if they planned any changes in the *next* five years, just over half—51 percent—said they did, mostly by expanding into criterion-referenced measures and the use of performance assessments.

Of those school administrators who have not changed in the past five years and have no plans to change, a solid 81 percent said they were reasonably satisfied with the outcomes of their current program.

Uses of testing

- 88 percent participate in a formal accountability program,
- 86 percent use their standardized test scores as part of the accountability program.
- Only 20 percent use those scores as the sole source of data.
- The other 80 percent report using several additional measures: district-made criterion-referenced tests (42 percent), student portfolios (31 percent), and "other" (40 percent).
- 91 percent use standardized tests in curriculum development and instructional supervision.

Cheating and pressures to improve

While the issue of cheating got its share of headlines in the past few years, only 5.4 percent responded positively when asked if their districts had any incidents of "cheating" by teachers or schools.

But 57 percent feel pressured by the public to improve results—and a surprising 48 percent feel this pressure is justified. When asked how they helped ease those pressures, there was a variety of answers:

- Only 5 percent said they lobbied for the repeal of mandated standardized testing.
- 17 percent developed and substituted criterion-referenced tests where allowed.
- 52 percent maintain a strong public relations program to explain the true significance of the tests.
- 37 percent conduct formal instruction of students on test-taking skills.
- 41 percent conduct inservice programs on how to improve their students' test scores.

Although 52 percent of the respondents said earlier that they employ a strong public relations effort, only 20 percent of the districts have developed any original materials to communicate with the public.

Comments on NAEP

The survey asked about the new direction taken by the National Assessment of Educational Progress (NAEP) in establishing student objectives and measuring progress. A surprising 65 percent felt that this would be "a welcome additional resource," while 35 percent called it "a dangerous trend."

The balance shifted, however, when it came to proposals to release NAEP test results on a comparative basis. State-by-state comparisons were opposed by 56 percent, while district-by-district and school-by-school comparisons each were called a "dangerous trend" by 71 percent.

Instructionally sound but biased

Of the respondents, only 27 percent are personally concerned that their district's efforts to improve standardized test scores may not be instructionally sound, but only 5 percent worry that they are morally or ethically improper.

More than half—57 percent—agree that standardized tests are biased against some racial and ethnic groups, but only 30 percent reported any gap in their own scores based on race, gender or ethnic group.

Are they important?

When asked to rate the importance of standardized test results in making policy decisions and developing curriculum:

- 31 percent responded that they and their school boards believed them to be very important.
- 63 percent felt they were slightly important.
- 5 percent saw them as irrelevant to what they were teaching students.
- 1.3 percent said they were counterproductive in meeting their instructional goals.

How are they doing?

And how are the schools doing overall? As if to add credence to the legendary "Lake Wobegon Effect" (where all the children are above average):

- 68 percent reported their scores to be above the national norm.
- 21 percent are at the national norm.
- 10 percent are below.

(More on this phenomenon later.)

What the Words Mean

A Concise Glossary of Assessment Terms

A

Accountability—The demand by public officials, employers and taxpayers for school officials to prove the educational impact of the money invested annually in education. This has led to the rise of what is called "accountability testing," designed to sample what large numbers of students have learned. This is contrasted to "instructional testing"—assessments designed to help teachers improve student learning in the classroom.

Achievement test—A test designed to measure a student's "school-taught" learning. It usually covers basic skills, such as reading, language arts and mathematics.

Alternative assessment—Generally, any kind of assessment technique other than traditional norm-referenced or criterion-referenced pencil-and-paper tests.

Aptitude test—A test intended to measure a student's innate ability to learn, given before receiving instruction.

Authentic assessment—A generic term for alternative assessment methods that test students' ability to solve problems or perform tasks under simulated "real life" situations.

C

Competency test—A test intended to make sure a student has met certain minimum standards of skills and knowledge and therefore should be promoted or graduated.

Criterion-referenced test—A test including questions based on what the student was taught to measure how much the student learned from that instruction. Can use multiple choice or essay questions, or combination of both.

Curriculum alignment—Altering curriculum to match the scope and sequence of a testing program, thus ensuring that teachers will cover material on the test.

Essay test—A test that requires students to answer questions in writing, emphasizing recall rather than choosing a correct alternative. The responses could be brief or extensive.

G

Grade equivalent—The grade level at which a student performs on a standardized test. A score of 5.5 means that the student is doing as well as the "average" student in the fifth month of the fifth grade.

H

High-stakes testing—Any testing program whose results have important consequences for students, schools, or school districts.

I

IQ (Intelligence Quotient) tests—The first of the standardized norm-referenced tests, developed more than a century ago, supposedly to measure a person's native intelligence. Used more today for psychological screening purposes.

Item analysis—Analyzing each item on a test to determine proportions of students selecting each answer. Can be used to diagnose particular strengths and weaknesses of students, as well as the test's validity or possible bias.

M

Multiple-choice test—A test consisting of items containing either a question or incomplete statement. Students are required to select the correct answer from the alternatives provided.

N

Norm—Distribution of scores obtained from the norm group. By definition, it is the midpoint of performance of the students in the norm group: 50 percent of them score above the norm and 50 percent below.

Norm group—A random group of students that originally take a test under the auspices of the test developer to determine the baseline and establish the percentiles of performance.

Norm-referenced test—An assessment instrument that relates the scores of each student to the scores of those in the norm group. A norm-referenced test tells how each student and group of students measure up against the original group that took the test, by distributing the range of scores along a continuum. These students do not necessarily fall half above the norm and half below. Only the norm group is so distributed.

Normal Curve Equivalent (NCE)—A normalized standard score with a mean of 50 and a standard deviation of 21.06. The score is most often used to enable the test user to manipulate the test data algebraically.

P

Percentile—Percentile ranks range from a low of 1 to a high of 99, with 50 denoting average performance. The percentile rank corresponding to a given score indicates the percentage of a reference group obtaining scores equal to or less than that score. For example, If a student scores at the 65th percentile, it means that he or she performed better on the test than 65 percent of the norm group.

Performance assessment—Also called performance-based, this type of assessment requires students to perform a task rather than simply answer questions. It may call for writing or solving math or science

problems or completing a science experiment. It is judged against established criteria.

Portfolio—A file of student work centered on a particular topic or content area. Now used as an assessment technique where specified contents of the portfolios are reviewed according to established criteria to determine the level of student performance and progress.

Project—A complex assignment that expects more than one type of activity and production for completion. A form of performance assessment.

Q

Quartile—After percentiles are determined, the distribution may be broken down for reporting purposes into four groups: the 0-25th percentile; 26-50, etc.

Quintile—A similar breakdown, but into five sections: 0-20, 21-40 etc.

R

Reliability—The measure of consistency for assessment instruments. A reliable test will yield similar scores when abilities or knowledge are similar.

S

Sampling—A way to get information about a large group by examining only a small number of the group (the sample), or by giving all members small segments of the test. When conducted properly, the results are considered highly reliable.

Standard—In assessment, a level of performance established by test users to differentiate scores obtained by students. The scores that determine the levels at which students "pass" or "fail," or are rated "satisfactory" or "unsatisfactory," are arbitrarily established according to the judgment of the testing authority. The word standard is also used in curriculum development to denote a level of achievement toward which all students should strive. The score at which they will be considered as having succeeded, however, is still an arbitrary judgment.

Stanine—Stanines, like percentile ranks, indicate a student's relative standing in a norm group. Stanines are normalized standard scores that range from a low of 1 to a high of 9, with 5 designating average performance.

Standardized test—An assessment instrument given to a large number of persons under similar conditions, designed to yield comparable scores. The term is most often applied to national norm-referenced tests developed by test publishers.

V

Validity—The measure of accuracy for assessment instruments. A valid test measures what we want it to measure, rather than extraneous variables.

T

Teacher-made test—A test prepared by the teacher for use in the classroom. It can be norm-referenced (designed to measure differences among the individuals in the class) or criterion-referenced (specifying minimum levels of acceptable performance), and performance-based. Since it is not standardized, the scores bear no relation to those obtained by students on tests given in other classrooms.

CHAPTER TWO

Issues and Concerns

It takes a lot of conviction to challenge an institution as firmly entrenched and widely used as standardized tests, but that hasn't stopped a growing band of critics from blaming them for many of education's ills.

Inside the administrative offices of one of the largest school districts on the east coast, there is a place called the "Agony Room." Here the superintendent meets regularly with his school principals. Permanently mounted on all four walls are poster-sized reproductions of each school's test scores, peppered with stars and flags highlighting scores that are "good" and those that are not.

When this superintendent was hired, he had promised the school board he would improve the district's below-average results on the state-mandated norm-referenced standardized tests. Thanks to the "Agony Room" and similar pressures, he succeeded and was acclaimed by the community. But, some critics ask, does that mean those students received a better education?

That question might seem frivolous to the 95 percent of the AASA survey respondents who administer norm-referenced standardized tests and the 94 percent who use them in making policy decisions and developing curriculum.

That level of popularity, however, is exactly what troubles the experts who have been filling the pages of professional journals recently to blame the nation's reliance on norm-referenced tests for a multitude of ills. Even some of the people who use them regularly recognize that the high-stakes nature of standardized testing programs raises them far beyond their intrinsic educational value, and said so on the AASA survey.

In response to a question where anonymity was offered in exchange for frankness, a testing director in a northwest school district said he was upset because his administration had trumpeted a new campaign to raise scores on a standardized battery as a indication of success. "This is a bankrupt, cheap focus on an inconsequential variance in test scores," he wrote. "I do not like it." And to show that the issue is not regional, a New England administrator commented, "I have no interest in improving our standardized test scores. It is not educationally defensible."

A coalition of dozens of national education groups calling itself the Campaign for Genuine Accountability in Education has gone on record as saying that "a heavy emphasis on standardized, multiple-choice tests as the primary measure for educational ability or achievement will undermine needed educational reforms in our nation."

The concerns

While there is some overlap, the most common complaints by critics sort themselves into five basic categories. The charges are that norm-referenced standardized tests:

- ***Are spurious***—that they give false information about how well a school is teaching or a child is learning.
- ***Are inconsequential***—that they focus on the simpler skills that are easily tested and take away from higher-order thinking skills and creative endeavors.
- ***Are counterproductive***—that they reduce teaching to mere preparation for testing.
- ***Are biased***—that they are unfair to some groups of students such as females, blacks, the limited-English speaking or those from low-income families.
- ***Are phony***—that you can't tell a real score from a bogus one because teachers are pressured to cheat to improve their classes' numbers. It was this charge that forced the assessment debate out of the educational journals and onto network television and the front page. And while many educators try to dismiss such incidents as isolated and overinflated, the publicity has colored the perception of the public and cast a shadow over the entire process.

Uses and misuses

"So long as test users were aware of the limitations of the information they received, these tests served schools well," according to Professor Herbert C. Rudman of Michigan State University, who wrote in the Fall, 1987 edition of

Educational Measurement. These were in "monitoring the achievement levels of individual pupils and the aggregate groupings of these students in terms of classrooms, buildings, and the district."

What has happened, however, is that some users have forgotten that these instruments test only a small segment of the total curriculum—a figure put by some critics at 30 percent.

"For the past decade or more," Rudman went on, "some test users seem to have lost sight of the fact that test content remains a *sample* of the total domain." In some cases, he said, the users have demanded that the test publishers add items that reflect their *total* curriculum. The result, he concluded, has been "inflated promises" by test publishers that have failed to meet the "unrealistic demands" of the users.

Thus, as a means of monitoring how a school is teaching the basic concepts of third-grade math, a norm-referenced standardized test can give a pretty good idea. And in the hands of an experienced testing specialist, an individual child's score can be a useful reference point to consider in analyzing progress and planning future instruction. It is when these scores are used for other purposes that their instructional value diminishes and the real problems appear.

Perceptions and pressure

One of the biggest problems a community can have is to see the development of a "scoreboard mentality" among its citizenry. This can turn the quest for quality education into a crusade for numbers. "I think the effects of this pressure are not well understood by the district, nor are the devised tactics always in the best interest of all involved," a local school district testing director in Florida said in his survey response when asked whether he felt pressure to increase scores in his district.

The dangers are magnified where test scores are released as part of a state or regional accountability program. Then it is not unusual to find newspapers ranking schools or even districts from "best" to "worst" based on their average scores on some standardized norm-referenced battery. It is also common to see real estate agents using test scores to promote one neighborhood over another.

In fact, several years ago pressure from local districts forced the state of Pennsylvania to abandon the practice of releasing—in rank order—district-by-district scores on its standardized testing program. The insurrection was led by the urban districts, which habitually ended up at the bottom of the list.

Even for those school systems that appear at the "top" of such a list, it is debatable whether test scores alone are a valid indicator of instructional quality. Yet the public perception that high scores mean quality education can put extreme pressure on school officials. Pressuring school staffs to raise scores may be educationally unsound, but it can become politically prudent.

Teaching to the test

The pressure to improve test scores can, in the long run, seriously hurt the imagined beneficiary—the child. But there can be a fine line between helping students feel comfortable with a test and inflating their scores.

Test-taking skills

According to several surveys, most educators believe that not only is it proper to teach to the domain of knowledge represented on the test, but it is the smart thing to do—as long as other curricular goals are not sacrificed. And teaching test-taking skills, they say, just makes sense. In fact, 37 percent of the respondents to the AASA survey said they offer formal instruction on test-taking skills to help ease the pressure to raise scores.

What's so bad about this? many people ask. After all, they argue, you don't send kids out to play a soccer game without first going over the rules and letting them run around the field. Or as one teacher ruefully commented, "If you're afraid of teaching to the test, just give your French class a Spanish test."

Crossing the line

While test-taking skills has become an acceptable unit taught in most schools, at what point does test preparation and practice, often using sample forms supplied by the publisher, cross the line and become not only cheating, but counterproductive to student learning?

Professor Lorrie A. Shepard of the University of Colorado at Boulder answered this way: "One-time practice with test format, especially when these activities are consistent with standardization procedures, is not the cause of inflated test scores. However, repeated practice or instruction geared to the format of the test rather than the content domain can increase scores without increasing achievement," she wrote in the Fall 1990 edition of *Educational Measurement.*

Teaching to a norm-referenced test, Shepard added, forces teachers to:

- Spend excessive time on those topics they know are going to be tested instead of the required curriculum.
- Emphasize drill and practice over essays to prepare their students for multiple-choice formats.
- Have their students memorize factual tidbits instead of

challenging them to develop their higher-order thinking skills.

There would be no question whether it was cheating if a teacher sneaked into the office a week early and made a Xerox copy of a test. But what about experienced teachers whose familiarity with the test over the years allows them to improve their students' performance? Shepard speculated this way:

"Suppose you couldn't help but remember several of the vocabulary items on the test, or you chose to do a science unit on one of the animals discussed in a test reading passage. Perhaps you were distressed during the test last year when your third graders were asked to do money problems in a format they had never seen before, so you decided to use examples in that format from now on."

Does it really matter?

Could this kind of familiarity make that much difference in the test scores? Shepard looked at the norms tables of two of the most popular tests, the California Achievement Test (CAT) and the Stanford Achievement Test (SAT), and concluded that one additional item correct translates into a gain of 2 to 7 percentiles.

The old complaint against standardized tests was that they were supposedly insensitive to instruction. "It would take an enormous amount of instruction aimed at the full domain to move the class average by a single item," Shepard concluded. "However, teaching to specific items is enormously more efficient."

Is 2 to 7 percent per item significant? Clearly. But is it cheating? And is the student really being short-changed in the final analysis?

It is perhaps not so important to answer those questions as to know that they are being asked. It represents one more reason for questioning whether the assessment vehicle that has served American education so well for so long should be overhauled or traded in for a new model.

Narrowing the curriculum

When standardized test scores are low and teachers feel pressured to raise them, the student can lose for another reason too. When the test comes from "outside," one answer is to sacrifice content.

In fact, Professor Shepard surveyed state testing directors for the Center for Research and Evaluation, Standards and Student Testing (CRESST) at UCLA, and learned that "narrowing the curriculum" was a very real explanation in many cases where improved standardized test scores were evident.

"Narrowing the curriculum is a standard and accepted practice," John Hollifield explained in the *R&D Preview* in describing Shepard's research. "States and districts select norm-referenced achievement tests that best match the curriculum. Committees of teachers are often appointed to make sure that such a match occurs. Face it. A state or district would have to be pretty dumb not to do this."

In her survey, Shepard noted that "once the test that best fits the curriculum is chosen, the practiced curriculum is adjusted further in response to the test." As Hollifield explained that phenomenon: "The curriculum is adjusted to make sure it covers the objectives covered on the test. In less pleasant words, tests drive the curriculum."

The "de-skilling" of teachers

Another outcome of this practice, Shepard wrote in the April 1989 edition of *Educational Leadership*, is that it drives out good teachers and "de-skills" those who remain.

"Teachers with the least content knowledge might feel secure in leading students through worksheets fashioned after the tests, even if they become less skilled as teachers in the process," she wrote. "Teachers who know the most about content, however, are insulted when external mandates prevent them from using their own expertise to devise instruction. They must either find a way to resist the de-skilling . . . capitulate, or get out of teaching."

Cheating

In the midst of the scholarly debates about the validity of standardized testing, the "C" word burst upon the educational scene in full force in 1988. Could it be true that educators were actually cheating to raise test scores?

This is not to say that word never came up in the past. As long as there have been tests there have been cheaters. But security was always a priority in standardized testing and such incidents were believed to be isolated and their impact minimal.

And then in 1988 John Cannell, a West Virginia physician, published a survey report that infuriated educators and intrigued editors. How is it possible, he asked, for all of the 50 states and 90 percent of their school districts to report that their students were testing above the national norm in achievement?

The "Lake Wobegon Effect"

It was a natural for the media, especially after the phenomenon was dubbed the "Lake Wobegon Effect" (for Garrison Keillor's mythical town where "all the men are strong, the women are good looking, and all the children are above average"). When Cannell answered his own question, could the TV cameras be far behind?

In a follow-up study, Cannell, who by then had moved to Albuquerque, New Mexico, to study psychiatry and formed a group called Friends for Education, concluded that a major reason for the uniformly high test scores was cheating. He solicited and received letters from teachers declaring that cheating was widespread, triggering investigations that turned up scores of incidents in several states.

The ensuing headlines in major newspapers all over the country and a segment on "60 Minutes" were not about the gray area of "teaching to the test." When they charged "cheating," they meant teachers giving answers to the kids or tampering with the answer sheets. The South Carolina teacher featured on the CBS program insisted she had no choice because the stakes were too high—job security and a cash bonus tied to raising her students' test scores.

How to Prevent Cheating

Not even the most vociferous critic of standardized tests suggests that they be eliminated entirely.

John J. Cannell, who probably gives test publishers a bigger headache than anybody, focuses most of his recommendations on increased security, enforcement of stricter guidelines in administering tests, and less pressure on the teachers.

Here are some of his recommendations:

- Give sealed test packets to teachers the morning of the test, or no later than the night before, rather than days in advance.
- Require teachers to turn in the tests immediately upon completion rather than days later.
- Hire proctors from outside the system to administer the tests (rather than continue to use the teachers who have a stake in the outcome).
- Have answer sheets checked by an independent agency for suspicious erasures.
- Change test questions significantly from year to year to prevent teachers from using unauthorized copies to "teach the test."
- Remove achievement scores from the criteria used in teacher evaluations and career advancement.

"Dumb luck"

Cannell had branded a lot of sacred cows with his accusations, so the U.S. Department of Education funded a study to come up with a technical explanation for the apparent paradox of the "Lake Wobegon Effect."

Gary W. Phillips of the National Center for Education Statistics summarized the findings this way: "They found that his survey methodology was questionable, that many of his arguments were fallacious, and that his major finding could actually be attributed to dumb luck."

The "dumb luck" was that all 50 states were indeed above the national average that year. Soon afterward, test publishers and education researchers were brought to Washington by the Department of Education, and agreed unanimously that such a phenomenon can happen for a lot of reasons other than cheating. None of those explanations, by the way, was calculated to catch the attention of a reporter or editor. And none did.

Still, Cannell's report drew attention to a situation that needed a closer look. "The report had uncovered a contradiction," Phillips said. "While local educators were preaching rigorous standards for accountability, they were practicing procedures that led to positive self-evaluations."

Scrutiny continues

Neither Cannell's Friends for Education, nor the issue, will go away soon. The picture, accurate or not, of institutionalized cheating in the nation's schools has caught the imagination of the nation's education writers, who are likely to pay increased attention to testing issues from now on. The respondents to the AASA survey, however, reported that incidents like those were few and far between as far as they were concerned. In fact, only 5.4 percent admitted a "cheating" incident in their districts and all considered them minor.

Still, even if the media don't get wind of it and the district is spared embarrassment, even a trivial episode can have long-lasting repercussions. A west coast testing director noted in his survey response that a "minor" incident in his urban school district had skewed some scores upward before the practice was stopped. "This will have an adverse affect for years to come," he said. Even though the higher scores were an aberration, those numbers will always be there for future comparisons.

A question of bias

Another major area of concern complicates the issue in many school districts. Do nationally normed standardized tests discriminate against groups of students because of

their gender, race, geographic location, socioeconomic status, or national origin?

A surprisingly high percentage (57 percent) of survey respondents felt this was true, but a much smaller number (30 percent) saw it as a problem in *their* districts. Many districts still do not "disaggregate" their data by subgroups, and thus could not answer the question. In other cases, where a gap in scores was evident, the respondents said they instituted remedial programs to improve the performance of those students with low scores, whatever their race or ethnic background.

That, of course, doesn't do anything about the tests themselves or really answer the question: are the tests biased?

The publishers insist they are going out of their way to eliminate bias from their tests, blaming the lower scores of minority students on the inequities that exist in American schools and American society. But critics of standardized tests concede that while these inequities do exist, the tests do not merely reflect them—they compound them.

"Standard" vs. "nonstandard" English

It is not merely a question of English-speakers versus non-English speakers. These critics point to sophistications in the language, and especially the stylized English commonly used in standardized tests, that prevent the accurate measurement of students who speak "nonstandard" English. This can be as real a problem to refugees from Asia and Central America as it is to U.S.-born African Americans, mountain families and the inner-city working class.

Similarly, there are other differences relating to cultural perspective and even geographic location that can similarly interfere with the valid measurement of a student's abilities. Some children may have had no experiences with a situation depicted in some questions and thus cannot relate to it like other children. Others may respond to a word or a situation in a manner unique to their culture—and not the way intended by the test maker.

A true story that made the rounds of testing directors in Maryland involved a question on the Wechsler Intelligence Scales for Children given in many schools. It asked:

"What is the thing to do when you cut your finger?"

The suggested answers (and their point values) were:

2-point response: "Put a Band-Aid on it."

1-point response: "Go to the doctor (hospital) . . . "Get it stitched up."

0-point response: "Suck blood . . . don't panic."

Minority children in urban areas usually perform poorly on this item. A sociologist from Baltimore, Maryland, asked inner-city youths why a large number of them answered "Go to the hospital." Their explanation was that in the vernacular of the streets, "cut" means a *big* cut. When the youngsters were told that the question referred to just a little cut, virtually all of them answered, "Put a Band-Aid on it."

Socioeconomic status

The issue, however, may not actually be one of race or ethnicity. Professors Norbert A. Stirzaker and Fredric L. Splittgerber, writing in the University of North Carolina's *Education Report* of November 1989, noted: "The school learning that took place traditionally in the house at mother's knee has disappeared because two incomes are frequently necessary to maintain a minimum standard of living."

The authors dismiss the various solutions offered in the testing controversy, such as renorming tests and raising standards, and conclude that none of them will reach the heart of the problem.

"Social institutions must seek better ways of overcoming socioeconomic factors that tend to impair the educational opportunities and learning experiences of children," they wrote. "Only parental expectations can improve the home environment. Learning cannot take place without home, community, and school cooperation. The dilemma is not how to improve test scores but how to establish equal educational opportunities."

A Nebraska testing director noted in his survey response that test scores in his district indeed show a decided gap when reported by race or ethnic group, but he disagreed that it was due to bias. "I think test scores rather accurately reflect opportunities that are available. The rich get richer," he said.

The three-year study by the National Commission on Testing and Public Policy, funded by the Ford Foundation, found the same phenomenon existing throughout the country, and put it succinctly:

"A significant segment of society . . . has experienced and continues to experience testing as a hostile gatekeeper."

Blaming the norms

Although at first glance, the "Lake Wobegon Effect"—where all the children are above average—suggests something rotten in the state of testing, many assessment experts blame the apparent contradiction on the outdated norms that in some cases compared achievement in the 1980s and 1990s against an anonymous norm group of the 70s. For whatever reasons, students did better and better on the tests, and since their performance was rated against an earlier generation, scores soared.

So what? many publishers responded. As long as every-

body is using the same norms, they argued, the comparisons are still valid. The problem arises when "above average" scores on the tests are interpreted as meaning the student, the district, or the school is "above average" in all educational measures.

But that's exactly what the media and the public insisted on doing, and many school districts did not bother to disillusion them. Still, they were in for some rocky days when it came time to adopt new norms. Virtually without exception, scores dipped sharply from the previous year. While this might tend to explain the Lake Wobegon Effect after all, it also creates a major public relations crisis.

Annual norms

One of the alternatives offered by test publishers is to employ annual norms. However, experts are skeptical. While annual norms might cure what ails Lake Wobegon, wouldn't they exacerbate many of the problems affecting real-life school districts?

For one thing, most districts use national norm-referenced tests because they are economical. A system of annual norms would negate that advantage because it would be exorbitantly expensive to establish and maintain. For another, this would create a "moving target" that would render comparisons and/or trends even more dubious than they were before.

Turning to "user norms" rather than a static "norm group" sounds like a valid alternative—if schools could afford the price tag—but it opens the door to other concerns. Shepard also addressed this issue in her research. For example, she pointed out, such norms would have to

"We don't produce a lousy car and wait for the consumer to judge it, yet we do that all the time with students."

—*Grant Wiggins*

be representative of the national population with no demographic biases. Nobody has yet confronted the question of what it would mean to adopt a concept of a norming population where everyone is teaching to the same test.

Ordinarily, a district is at a disadvantage if it plays fair by teaching to the broad curriculum and avoiding excessive practice on test format. "The disadvantage would be exaggerated in a comparison among users rather than one in which each is compared with the naive norming sample," she added. Since there is no way to assure that everybody in the norm group will teach to the test in the same way, their invalid gains would be built into the norm. "The standard of comparison based on user norms would be spurious and inflationary," she concluded.

Thinking skills sacrificed

Much of the concern about standardized norm-referenced as well as criterion-referenced tests is that they emphasize rote memorization at the expense of thinking skills and creativity. These criticisms are based as much on the traditional format of such tests as their content. The problem, they say, is that as long as students respond by selecting a correct answer that someone else wrote, we may never know if they can actually use their skills and knowledge in any realistic, life-like setting.

"Bubble-bashing"

There is a growing attitude on the part of many educators today that "if it's multiple-choice it must be recall." Jason Millman of Cornell University dubbed it as "bubble bashing," summarizing the alleged weaknesses of multiple-choice tests as follows:

- They test recall of facts, not thinking skills.
- Preparing a class for multiple-choice tests promotes poor teaching.
- The tests cannot be used to predict future performance in class or beyond.
- They don't measure many important traits like persistence and the ability to organize time and work.
- They are biased against minorities, the poor, and women.

Even the revered Scholastic Aptitude Test came under attack during this period, in large measure because of its continued reliance on multiple-choice questions. In response, *The College Board Review* published an article late in 1990, which it reprinted in brochure format, defending multiple-choice testing against its critics.

"Many multiple-choice tests are more than *just* tests of knowledge," the article said, contending that the SAT "tests very little knowledge; it is almost completely a test of analytical and reasoning ability at quite complex and sophisticated levels. The multiple-choice questions on Advance Placement Exams—about half of each test—often require sharp analysis with a command of facts."

The author, Lawrence Feinberg, a long-time education reporter for The Washington Post who later became assistant director for reporting and analysis of the National Assessment Governing Board, conceded that some multiple-choice tests are "simple," citing as an example minimum competency tests. But he said they are made easy for "political" rather than educational reasons. An exam's level of difficulty, he argued, is not a function of its format.

Nevertheless, the pressure was growing and in October 1990, the College Board's trustees recommended that revised and expanded versions of the SAT verbal and math-

ematics tests (SAT-1) include increased emphasis on critical reading and require student-generated answers in math. There will still, however, be multiple-choice items mixed in with the open-ended ones.

A question of validity

Test format aside, this leads to another critical concern about traditional standardized tests—that of their basic validity. Do they really test what we want them to test? Do they cover the depth and breadth of the curriculum, or do they skim a little off the top? As we will discuss in more detail later, the state of Arizona launched in 1990 a major overhaul of both curriculum and assessment after a survey showed that the existing norm-referenced instruments covered only 26 percent of the state's curriculum.

And another key question: Do they test the full abilities of the student? There has been considerable progress lately in curriculum and instruction in developing ways of challenging students' higher-order thinking skills, but many observers feel that the teaching has gotten way ahead of the testing.

The goal of our schools is not to have students restate facts we have given them, Richard Stiggins reminds us. Rather, "Our goal is to teach students to use the information at their disposal. We want them to think." How, then, do we assess thinking? Researchers like Stiggins have devoted a great deal of attention to developing assessment methods that emphasize measuring students' thinking skills. These will be discussed in Chapter 7.

Standards, not scores

At the root of much of the controversy over assessment is the conviction that the way to improve education is to set standards and develop uniform ways to measure them. But, they feel, as long as schools insist on doing business as usual it's not going to happen. Dale Carlson, director of the pacesetting California Assessment Program, decried the "insidious effects of standardized testing—in which we focus on scores, not standards"—as putting the United States "40 years behind the rest of the world" in assessment.

Lauren Resnick, a respected researcher in the field of assessment who directs the Learning Research and Development Center at the University of Pittsburgh, agrees. "The United States is the only major nation in the world in which students go to school without working toward an examination standard," she said. "In other countries, all students—not just those going to college—are expected to meet a high standard of academic achievement and to take a curriculum that will prepare them to do so."

Such examinations, she said, if tied to educational reform that provides a clear message of the desired outcomes, might change society's belief "that aptitude and family determine success in school" to one where "what you do in school makes a difference."

The debate that began in 1990 over proposed national standards, national goals and national tests reflected that same concern. Colorado Governor Roy Romer, chair of the National Education Goals Panel (NAGB), explained that in his mind, goals and assessment can improve education in two ways: by setting benchmarks that teachers, students and schools can aspire to, and by giving a more accurate portrayal of student skills and achievement.

The national debate is covered further in Chapter 8. But meanwhile, what will become of the controversy over standardized testing in the nation's school districts?

Looking ahead

The Campaign for Genuine Accountability in Education has pulled no punches in lobbying political and education leaders at all levels to phase out most uses of standardized multiple-choice tests. These leaders, the statement says, "should not rely on existing standardized tests as indicators of educational progress; and should not depend on standardized tests to make decisions about program placements for students. These tests misdirect the curriculum, are inadequate educational measures, are not instructionally useful, are often too inaccurate to be the basis for making decisions about individuals, and can be replaced with authentic assessment instruments."

The gauntlet is thus thrown, but the battle promises to be a long one. The American educational system has too much invested in national standardized testing to give it up easily. Many users insist that these tests serve a valid purpose, and they do it at far less cost than any alternative in-house effort.

Publishers are responding to the challenge by updating norms more frequently, providing more opportunity for curriculum alignment, including items that require students to apply higher-order thinking skills, and offering a greater variety of test items from year to year.

There are, as the statement suggests, many movements under way to replace these tests with alternative forms of assessment. What is still not certain is whether these techniques can—or ever will—come up with the kind of data that school boards and the public will demand.

No less an authority than Dale Carlson, who pioneered authentic assessments as director of the California Assessment Program, says the real challenge is whether they can also provide hard measures of two vital but very different goals—school quality and accountability. "We must have the right kinds of assessment," Carlson said, "but we also must have the hard numbers."

CHAPTER THREE

The Search for Alternatives

Under pressure from many directions to clean up their assessment act, the educational community has set out to explore new and different ways of measuring the success of the schools and their students.

When testing reached its "crossroads" in the early 1990s, one of the choices, to continue the metaphor, was a narrow mountain road. On one side was a steep cliff representing increasing pressure to document the quality of education. On the other side was a deepening abyss of meaningless numbers. And although the search for alternatives was uphill, there was no turning back.

There had been the other choice—to stay on level ground pretty much as before—but as we have just seen, the institution of testing was taking increasing heat for failing to measure accurately—and to improve—student learning. It could be said that the institution was in a rut.

"Testing was developed in the first place to serve individuals who were not directly involved in instruction and learning," observed Pittsburgh's evaluation director, Paul G. LeMahieu. "Their work was at some distance from the classroom. They were what we refer to as decision makers, policy makers, etc." Just about 100 years of testing have gone by, LeMahieu continued. For the first 80 years, "we believed that it was possible to develop tests that were neutral in regard to curriculum and instruction." For the last 20 years, he added, "we have known better, but unfortunately we haven't been honest enough to behave any differently."

It would have been impossible, however, for educational testing to stay in its 100-year-old rut in the face of such criticism as that leveled by the National Commission on Testing and Public Policy in May 1990. Its report, "*From Gatekeeper to Gateway: Transforming Testing in America*," used strong terms in denouncing the current reliance on standardized multiple-choice examinations (both norm- and criterion-referenced).

The panel, composed of leaders from education, business, labor and civil rights groups, called for a transformation of the current American testing system. It urged schools and businesses to use alternative forms of assessment that would measure performance in school and on the job, and to consider a broad range of information—including but not limited to test scores—to evaluate individuals and institutions.

That kind of report was becoming a trend.

A citizens group of similar makeup had been appointed two years earlier in Maryland by newly elected Governor William Donald Schaefer. He had campaigned as a supporter of education, and now wondered how employers could be complaining about the skills of high school graduates when the state's standardized test scores were so high.

The Governor's Commission on School Performance zeroed right in on the state's norm-referenced testing program. In August 1989, it urged in its final report among other things the adoption of a statewide accountability model based on multiple indicators of school health and an alternative method of testing student achievement that:

- Measures fundamental, important things in the curriculum.
- Measures more than a narrow, low-level set of objectives.
- Does not encourage rote teaching of material covered by the test.
- Holds schools, in addition to students, accountable for achievement.

The panel's recommendations received immediate and unqualified support from Governor Schaefer, and within months, Maryland's schools had abandoned the norm-referenced tests and were busy developing new, more real-life measures of reading, writing, mathematics, science, and social studies. Maryland thus joined a small band of pioneers at the state and local levels in the search for alternative assessments that would both reflect and influence learning in the classroom.

Alternative to what?

When the phrase "alternative assessments" became popular in the 1980s, it was easier to explain what it did *not* mean than what it did. Alternative assessments were many

A Vision of Educational Assessment in California

What If . . .

The following "vision statement" addresses most of the issues facing educators today as they work toward developing the "ideal" testing program. It was written by director Dale Carlson and his staff of the California Assessment Program (CAP) in 1989 to motivate the participants in statewide Curriculum/Assessment Alignment Conferences called "Beyond the Bubble." It is reprinted here with permission.

What if
students found assessment to be a lively, active, exciting experience?

What if
they could see clearly what was expected of them and believed that the assessment provided a fair opportunity to show what they had learned?

What if
they were challenged to construct responses that conveyed the best of what they had learned—to decide what to present and how to present it—whether through speech, writing or performance?

What if
they were educated to assess themselves, to become accurate evaluators of the strengths and weaknesses of their own work, and to prescribe for themselves the efforts they must make to improve it—ultimately, the most important form of assessment available to our students?

What if
the assessment allowed students to use their own backgrounds and indicated ways of building on their strengths for further learning?

What if
their learning were recognized by the school and the community when they had made outstanding progress—regardless of their initial level of achievement?

What if
teachers could look at the tests and say, "Now we're talking about a fair assessment of my teaching;

one that
focuses on the essence of the student outcomes that I am striving for—not one that focuses on the peripheral skills or the isolated facts which are easiest to measure;

one that
shows they can produce something of value to themselves and to others—an argument, a report, a plan, an answer, a solution; a story, a poem, a drawing, a sculpture, or a performance; that they can conduct an experiment, deliver a persuasive oral presentation, participate cooperatively and productively in groups;

one that
is accessible to all my students, yet stretches the most capable students as well—not one that measures some mythical minimum competency level;

one that
matches the assessment that I use on a day-to-day basis to guide my teaching and that guides my students in their learning—not one that takes an artificial form and then naively expects students to give a natural response in an artificial situation; one that doesn't take valuable time from the teaching/learning process, but is an integral part of that process;

one that
doesn't treat me as a 'Teller of Facts,' providing and prescribing the concepts and the content that students are to study—but rather as a coach and fellow learner, helping my students to become active learners who are prepared to discover what is important to them now and enthusiastic about learning in the future?"

What if
the new assessments led parents, taxpayers, legislators, and the business community to exclaim:

"I can see
that the schools are focusing on the important things—that students are achieving levels of academic excellence which truly prepare them for the future;

I can see
that students are learning what they need to fulfill themselves as individuals, to become concerned and involved citizens and workers who can adapt to the changing demands of our world—creative people who can think and take initiative, who care about what they do, and who can work with others to solve problems;

I can see
the results in the newspapers which give me the information I need in terms I can understand, which show me the progress our schools are making on assessments that really matter, and where they need my help and support?"

The California Assessment Program
is committed to making this vision a reality;

a reality
for students who are being challenged to achieve excellence in schooling that truly prepares them for life;

a reality
for teachers who will find their best teaching assessed and rewarded;

a reality
for the parents, taxpayers, legislators and businesses so vitally interested in the success of the educational enterprise.

Please join us
in accepting the challenge to create the new forms of assessment which will make this vision a reality.

things, but primarily they were *not* multiple-choice, closed-response tests. In other words, they were *not* norm-referenced standardized tests, nor were they criterion-referenced tests that used the traditional multiple-choice or true-false format.

The movement away from relying on traditional tests received an early push on a building level from several members of the Coalition of Essential Schools, an effort launched by Theodore Sizer in the 1980s. They turned to alternative testing methods based upon the "exhibition of mastery" principle. Schools like Thayer, New Hampshire, High School have been singled out for their prototype system of "diploma by exhibition." As early as 1971, Walden III High School in Racine, Wisconsin, began using what it calls a Rite of Passage Experience (ROPE), in which students must demonstrate mastery in 15 areas of knowledge and competence by completing a portfolio of work and defending it before a committee made up of staff, a student, and an adult from the community.

Traditionally, teachers have had students in subjects like art, music, and physical education demonstrate their skills to get a grade. One of the earliest examples of this assessment technique on a widespread basis appeared in the NAEP exams in the 1970s. In art and music, students were asked to draw and sing; in math and science, they were asked open-ended questions. The conclusion was that this technique worked, but also that it was very expensive, so it was dropped from later NAEP assessments.

This concept, now called "performance assessment" or "performance-based assessment," has been resurrected as one of the most promising of the alternative techniques. Others are projects, a type of performance, and portfolios.

"Performance assessment" defined

In the face of legislative mandates and other pressures for accountability, many states and school districts began looking seriously in the 1980s at assessing performances.

Tests of this type, also called "authentic" assessment because they more closely resembled challenges of the "real world," differed from traditional assessment techniques in that they required students to demonstrate what they know, can do, or were taught, instead of just selecting an answer somebody else had written.

Richard J. Stiggins of the Northwest Regional Education Laboratory, one of the early advocates of this technique, defines it as containing four components:

1. A reason for the assessment.
2. A particular performance to be evaluated.
3. Exercises that elicit that performance.
4. Systematic rating procedures.

It doesn't really matter, in Stiggins' definition, where this happens. The open-ended demonstration could take place within everyday classroom activities or as a response to a carefully structured situation presented by a specially trained test administrator. The response could be given in writing or verbally, or even sung. It could require simple or elaborate apparatus or none at all, and students could work alone or in groups. The performance may be observed and graded on the spot, or may be recorded on video or audio tape for later scoring.

The major promise of performance assessments is that they facilitate improvement in instruction and learning, primarily because they conform more closely to important instructional goals than the traditional multiple-choice items on standardized tests, according to Robert L. Linn, professor of education at the University of Colorado at Boulder. "Indeed, the major promise of performance assessments is that they are so congruent with educational goals and good instructional practice that they will enhance instruction and facilitate the achievement of important educational goals," Linn said.

Performance measures

Edward D. Roeber, supervisor of the Educational Assessment Program for the Michigan Department of Education, offered a list of possible performance measures that have been used or proposed for use in various programs:

- Art: Draw, paint, participate in various art activities.
- Career Development: Apply for a job, interview for a job, participate in a work team to accomplish various work tasks, lead and follow in a work team.
- Mathematics: Measure, use calculators and computers.
- Music: Sing, dance, play a musical instrument.
- Reading: Use reference materials, read in free time.
- Writing: Write an essay or letter, write for enjoyment.
- Speaking/Listening: Speak in public, analyze conversations, communicate non-verbally, participate in a drama presentation.

But is that enough?

"These examples provide some of the most important outcomes that we hope our high school graduates take with them as they exit our schools," Roeber said. "But what evidence do we have that they actually do have these skills? Classroom teachers may be able to gather evidence that students have learned these skills, both through observation in the course of classroom activities and the classroom-level examinations they use. However, are these skills assessed and reported in a manner comparable across districts that the public, at the local and state levels, can use to determine student performance? Probably not."

Performances in years past—although they weren't called

But Is It Authentic?

Grant Wiggins, a senior associate with the National Center on Education and the Economy, has been on the cutting edge of assessment reform throughout the past decade. He has condemned traditional testing for "bypassing context-sensitive human judgments of human abilities in the name of statistical accuracy and economy." But he has also cautioned that performance assessments could be just as irrelevant and harmful to the educational process unless educators ask themselves the most basic question: "What is a true test?"

Wiggins asked that question, and answered it, in an article in the May 1989 *Phi Delta Kappan*. At that time, he labeled his own answer as "radical." The fact that it no longer seems radical proves how rapidly the assessment evolution—or is it revolution?—has been moving.

A true test of ability, Wiggins responded, is the performance of a task in a situation that most closely matches the standards and challenges of real life. And he introduced another term for the assessment vernacular: "authentic assessment." And while some users consider that phrase as a generic term interchangeable with "alternative" or "performance-based," Wiggins sees a subtle difference.

The task in the classroom, according to Wiggins, is to recreate the context in which problem solving, critical thinking and creativity occur naturally.

He writes that authentic assessments must first "replicate the challenges and standards of performance that typically face writers, businesspeople, scientists, community leaders, designers, or historians." Students can respond to the "questions" by writing essays and reports, conducting interviews, designing proposals, compiling portfolios, and so forth.

Furthermore, he added, an authentic assessment must be sensitive to individual students as well as to the individual school context. "Evaluation is most accurate and equitable when it entails human judgment and dialogue, so that the person tested can ask for clarification of questions and explain his or her answers."

that—were primarily activities designed to culminate classroom units. Their "grading," if there was any, was almost entirely subjective. The outcome of the "performance," or group project, never appeared in the school's achievement profile. The major chunk of each student's course grade still came from pencil-and-paper exams and the school profile came from standardized tests.

Thus, the pioneers in the alternative assessment movement faced two major challenges:

- To design performance assessments that would truly reflect the full domain of the curriculum and help teachers determine what instructional strategies should be used for each student.
- To figure out how to score those performances to meet the demand for accountability by giving a valid picture of the achievement of each student, school, and school district.

In other words, while the word is generally shunned because of its historical connotation, alternative assessments had to become *standardized*.

States take the lead

For the most part, state departments of education have been taking the lead in the movement to replace standardized multiple-choice tests as the major ingredient in their accountability formulas.

While there is no clearing house keeping track of school district assessment programs, a unscientific survey conducted in 1990 by two school district evaluation directors, Walter D. Hathaway of Portland, Oregon, and Joe B. Hansen of Colorado Springs, Colorado, indicated that about 100 individual school districts were exploring alternative methods of assessment, mostly performance-based. Many of the survey respondents found these activities useful in monitoring the effectiveness of instruction and in pinpointing the need for changes, but they were not designed to be used in any kind of large-scale accountability effort.

The writing assessments under way in more than half the states are also "performance assessments" used to determine students' skills. The assessments themselves range from very structured—such as the Maryland Writing Test—to others that do not even describe the topic that each student must write about. These programs fall more into the category of minimum competency testing used to ascertain whether students deserve a diploma, and thus they cannot be used to test the health of the educational system.

Michigan a pioneer

Michigan is generally credited with being the first state to add performance-based tasks to its achievement testing program. Such assessments were administered in 1974 to a statewide sampling of students in mathematics, music, science, social studies, art, music, career development and physical education. In most cases, the collection of data and scoring were done by volunteers to keep the cost down.

At least, said Roeber, who directed the Michigan Educational Assessment Program (MEAP) in the 70s, his state was able to demonstrate the feasibility of performance assess-

ment, and encouraged other states to try it on their own. While budget constraints forced MEAP to pull in its horns and become no more than a modest sampling effort, Roeber saw a revival brewing in the new atmosphere of the 90s.

East and west

It wasn't until a decade later that Connecticut incorporated performance testing in virtually every subject area—art and music, business and office education, English language arts, foreign languages, industrial arts, mathematics, and science—leading to across-the-board achievement standards.

New York moved ahead in performance evaluation after a successful program starting in 1988 in science. More than 200,000 fourth graders moved from station to station to show how well they understood the scientific process and whether they could use the laboratory apparatus.

California implemented sweeping changes in its statewide assessment program, involving hundreds of thousands of students. With a good two-year track record in writing assessment, California added mathematics and field-tested in other areas, such as language arts, science, history, and social studies.

Finally, three other states—Arizona, Maryland and Kentucky—began incorporating performance assessments with other statewide educational reforms starting with some preliminary activities in 1990-91.

Portfolios—another alternative

Vermont burst upon the scene by piloting a huge statewide effort in the 1990-91 school year. This one used performance tasks as a part of a comprehensive assessment program, but its real focus was on the use of student portfolios.

The word "portfolio" is a common one, especially in such subjects as art. In classrooms throughout the nation, many teachers routinely collect folders full of student work in many subjects to compare throughout the school year, or even from year to year. What's new in the 90's is the idea of using portfolios as the keystone of districtwide and even statewide assessment programs.

As the accountability movement began to grow, individual schools turned to the collection of student work as a more formal way of keeping track of achievement and to have something tangible to show parents to monitor their children's progress or lack of it.

In fact, 31 percent of the respondents surveyed for this book said they employed student portfolios in addition to norm-referenced tests in their accountability programs. In most cases, these are used strictly on a local school or district level. While they serve as an excellent means for teachers and parents to review student progress—especially in areas where tests fall short, their use has not yet been structured enough to become a true accountability measure.

"We didn't just sit down and write new tests. We started with the standards that kids ought to be able to meet, and based our tests on them. If that forces schools to revise their curriculum so they teach to our test, we've done our job."

—Arizona's Lois Easton.

Portfolios, however, may turn out to be the vehicle that carries the banner of alternative assessments over the long haul. Even though they are usually defined as one alternative form of assessment, in reality they call for a combination of assessments to make a better-informed judgment on the progress of students.

Higher-order multiple-choice

Many educators are not ready yet to give up on multiple-choice tests because of concern over high-order thinking skills. In fact, traditionalists maintain that multiple-choice examinations can and do tap higher order thinking. Anyone who has suffered through a *Miller's Analogies* test will have to agree. Here is an example from *Measuring Thinking in the Classroom* from the Northwest Regional Education Laboratory that is definitely *not* a rote recall item:

Was infantry invasion of Japan a viable alternative to the use of the atomic bomb to end World War II? If so, why? If not, why not?

A. Yes; transport ships were available in sufficient numbers.

B. Yes; island defenses in Japan were minimal.

C. No; estimated casualties would have been much greater.

D. No; Japan was on the verge of having an atomic bomb.

Here is what looks like a simple mathematics multiple-choice question submitted by Paul Williams of CTB/McGraw-Hill, but one that he says exemplifies the application of higher-order thinking skills:

A company will use vans to carry employees to a

What Is Intelligence?

Another recent entry into the assessment vernacular is a phrase originated by Harvard University Professor Howard Gardner: "intelligence-fair assessments." Gardner argues that a human being actually has seven intelligences: logical/mathematical, linguistic, musical, spatial, bodily/kinesthetic, interpersonal, and intrapersonal. Traditional assessments, he contends, may test only one of them at a time, and thus he offers another way of approaching the challenge:

"Intelligence-fair assessments are a useful alternative to paper-and-pencil tests.... For example, in contrast to a word problem, a drawing or painting task would be an 'intelligence-fair' assessment of Picasso's spatial skills. In order to assess the late dancer/choreographer Martha Graham's bodily-kinesthetic abilities, she would have to be allowed to leave her desk and dance; in order to consider Gandhi's interpersonal abilities, it would be necessary to see the effects and influence he had directly on other people."

Whether this kind of assessment could actually find its way into a school's test battery is not the issue, but it leads to another point about alternative assessment techniques in general: Rarely, if ever, is there one right answer. There are often multiple right answers and multiple wrong answers. Reliance on only one right answer is one of the artifacts of a purely multiple-choice format.

Thus the education community is faced with a special challenge as it moves away from the true-or-false, right-or-wrong world of the multiple-choice test and into the era of performance assessments: If our students are to leave school with skills that will benefit them in a variety of workplaces and in their personal lives, we must encourage them to attempt more complex tasks, demonstrate the strategies and thinking they use, and produce results that are not carbon copies of "right" answers.

This puts a greater burden on assessment specialists to develop scoring procedures that are flexible enough to give the student the benefit of the doubt. Machine-scorable tests may be quick and easy—and even cheap—but if they sell the student short, they have no place in education.

company picnic. Each van can carry 8 people. How many vans are needed to carry 36 employees to the picnic?

A. 3
B. 4
C. 5
D. 6

The answer in this case is C. As Williams noted, "The item requires students to analyze information and interpret a numerical response as it relates to a practical situation. Students must follow an appropriate problem-solving strategy and then deal with an unexpected result."

Items like these are being developed by test publishers for future norm-referenced instruments.

Norm's not dead yet

The emergence of performance assessment did not necessarily signal the demise of the norm-referenced standardized test. Most states that pioneered performance-based assessment retained a modest norm-referenced component at the same time. They either reduced the number of grades tested or the students sampled below pre-alternative levels, but the norm-referenced scores would still be compiled and reported as a sort of check and balance.

As described earlier, there are still valid educational uses for such instruments, especially with the addition of items calling for higher-order thinking. Besides, in spite of protests that comparisons of scores on norm-referenced tests can be highly misleading, nationally standardized tests remained the only non-subjective way for educators in one state to see how they shape up against others. Whether that's a valid reason or not, it was one of the rationales adopted by proponents of a national test.

There is one other explanation for the staying power of the norm-referenced test. It is a phenomenon evident in the responses to the AASA survey on assessment: Even where performance assessment was mandated statewide, in many cases the local superintendents failed to check that box on the questionnaire. But they did check the norm-referenced tests their district administers.

Michigan's Roeber, who keeps track of assessment issues nationwide, theorized that it boiled down to a question of "ownership." Local superintendents look upon state and federal tests as being "imposed" on them from outside, he said. "It will be a long time before many local superintendents claim such programs as their own."

"We must have the right kinds of assessment, but we also must have the hard numbers."
—California's Dale Carlson

Never the twain shall meet:

Maintaining Balance in Pittsburgh

Assessment seems very simple to Paul G. LeMahieu. All one has to do is distinguish between the two main reasons for testing . . .

- The evaluative: To learn if some person or some group, school, or school system is competent in something and thus get information for accountability reports to the public.
- The clinical: To diagnose what each child knows or does not know and to find the best way to teach him or her.

. . . and then give tests for one purpose or the other. *But not the same tests.*

"The same testing approach cannot serve both goals equally well," said LeMahieu, director of research, testing and evaluation of the Pittsburgh, Pennsylvania, Public Schools. He added what to him is an axiom: "The better job a testing program does with one, the worse it does with the other."

The key is to build and manage an overall assessment program that maintains the proper balance between the two, meaning not only the time spent on each part but the influence it exerts on decision making. In Pittsburgh, the ratio is 80-20:

- 80 percent of the effort goes into the home-grown programs called Monitoring Achievement in Pittsburgh (MAP), and the performance-based Syllabus Examination Project (SEP) for all secondary academic courses. Each uses both performance assessments and criterion-referenced tests to give frequent and immediate feedback to classroom teachers about each student.
- The other 20 percent goes into norm-referenced standardized tests for comparative purposes.

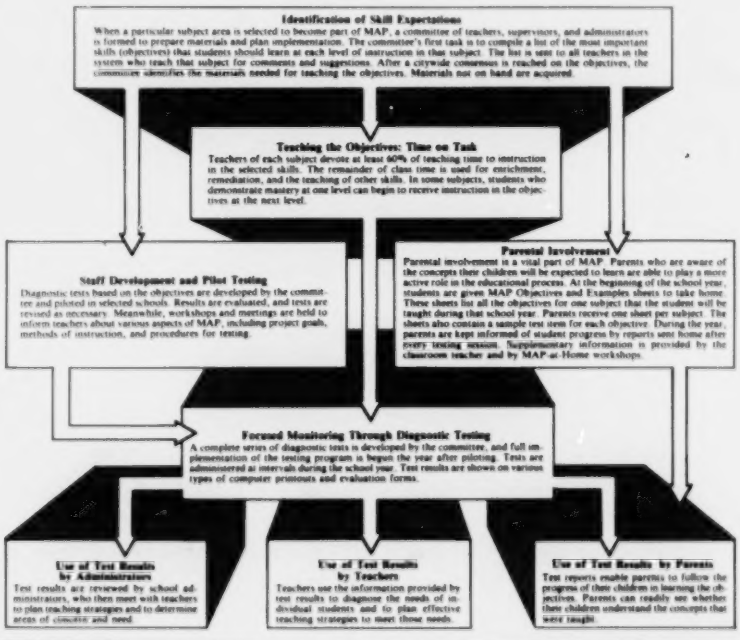
"Most districts have never tried seriously enough to develop tests for clinical and diagnostic uses," LeMahieu said, but even those that do eventually succumb. "Now that we have our test," they say, "let's also use it to compare our schools."

The MAP program is strictly diagnostic. It is used to help teachers plan strategies to meet the varying instructional needs of each student. The evaluations are computerized so that teachers get the results as quickly as possible. This is the only purpose for the tests. It is up to the teachers to make the critical decisions about what materials and strategies to use to help their students learn. Students learn because of, not in spite of, the efforts of the classroom teacher. As far as LeMahieu is concerned, one of the keys to his program is enhancing the teachers' self-esteem as a necessary step in improving their teaching skills. "I will always defend the teacher's right to exercise judgment, even if I might have done it a little differently."

How does such a program fly in the community? "Public pressure is less of a problem with two types of tests. When there is only one test, it gets to be high-stakes," LeMahieu said. "As long as we tell the public that we are using our program to diagnose kids' needs, they are happy."

The difference between the two types of tests is like night and day, LeMahieu said: "A child who doesn't understand something is supposed to tell the system. The most useful information we get out of a test is when a kid gets an answer wrong. Dare we say that in a political environment where a high score is important?"

Steps in the MAP Process


Identification of Skill Expectations
When a particular subject area is selected to become part of MAP, a committee of teachers, supervisors, and administrators is formed to prepare materials and plan implementation. The committee's first task is to compile a list of the most important skills (objectives) that students should learn at each level of instruction in that subject. The list is sent to all teachers in the system who teach that subject for comments and suggestions. After a citywide consensus is reached on the objectives, the committee identifies the materials needed for teaching the objectives. Materials not on hand are acquired.
Teaching the Objectives: Time on Task
Teachers of each subject devote at least 60% of teaching time to instruction in the selected skills. The remainder of class time is used for enrichment, remediation, and the teaching of other skills. In some subjects, students who demonstrate mastery at one level can begin to receive instruction in the objectives at the next level.
Staff Development and Pilot Testing
Diagnostic tests based on the objectives are developed by the committee and piloted in selected schools. Results are evaluated, and tests are revised as necessary. Meanwhile, workshops and meetings are held to inform teachers about various aspects of MAP, including project goals, methods of instruction, and procedures for testing.
Parental Involvement
Parental involvement is a vital part of MAP. Parents who are aware of the concepts their children will be expected to learn are able to play a more active role in the educational process. At the beginning of the school year, students are given MAP Objectives and Examples sheets to take home. These sheets list all the objectives for one subject that the student will be taught during that school year. Parents receive one sheet per subject. The sheets also contain a sample test item for each objective. During the year, parents are kept informed of student progress by reports sent home after every testing session. Supplementary information is provided by the classroom teacher and by MAP-at-Home workshops.
Formal Monitoring Through Diagnostic Testing
A complete series of diagnostic tests is developed by the committee, and full implementation of the testing program is begun the year after piloting. Tests are administered at intervals during the school year. Test results are shown on various types of computer printouts and evaluation forms.
Use of Test Results by Administrators
Test results are reviewed by school administrators, who then meet with teachers to plan teaching strategies and to determine areas of concern and need.
Use of Test Results by Teachers
Teachers use the information provided by test results to diagnose the needs of individual students and to plan effective teaching strategies to meet those needs.
Use of Test Results by Parents
Test reports enable parents to follow the progress of their children in learning the objectives. Parents can readily see whether their children understand the concepts that were taught.

BEST COPY AVAILABLE

Performance Assessment: Living Up to Expectations

Performance assessments show tremendous promise in transforming testing and improving testing and learning. However, one of the nation's leading experts on assessment warns that criteria for evaluating them are crucial to prevent them from becoming just another educational fad.

The following is excerpted from an article by Robert L. Linn, professor of education and co-director of the Center for Research on Evaluation, Standards, and Student Testing at the University of Colorado at Boulder, in *Policy Briefs*, a publication of the North Central Regional Education Laboratory.

Assessment policies and procedures at the local, state, and national levels are in a period of rapid transformation. The demands for greater accountability and higher standards are increasing the salience of assessment at the same time that there is a growing dissatisfaction with traditional methods of assessment and increasing concern about the unintended consequences of high-stakes testing programs. The resulting ferment has set the stage for major changes in assessment policies and practice.

Performance assessments . . . provide the vision that is guiding many of the efforts to transform assessment. Possibly the most important promise of performance assessments is that they will facilitate improvement in instruction and learning. Performance assessments are designed to engage students in solving problems and performing substantial tasks of importance in their own right. . . . If performance assessments realize their promise and lead to better educational consequences, they certainly will be worth the substantial effort that the construction and implementation of such assessments will require.

Of course, performance assessments are not all alike. They will not all be equally valid measures of important learning outcomes nor will they all have equally salutary effects on learning and instruction. Too often there is a great gap between intentions and implementation. Hence, there is a need to be able to determine not only the degree to which expectations are realized but also the relative validity and the nature of the consequences, both intended and unintended, of different performance assessments. In short, we need criteria for evaluating performance assessment.

Criteria for evaluating performance assessments

Traditional criteria, the most important of which are validity and reliability, do exist for judging measurements. These criteria, although relevant to performance assessments, must be thought of in broader terms than they normally are when applied in standardized measures. In *Complex, Performance-Based Assessment: Expectations and Validation Criteria*, Eva Baker, Stephen Dunbar, and I [Robert Linn] describe ways in which the criteria need to be expanded and applied to the valuation of performance assessments....

Consequences

The concept of validity has been expanded in the past few years to include concerns about the intended and unintended consequences of measurement. This expanded notion is particularly relevant to assessments, either traditional or direct performance assessments, that are intended to have an impact on student learning. The major promise needs to be verified. Evidence must be gathered on a systematic basis regarding both the unintended and intended effects of assessment on the ways in which teachers and students spend their time. It is not enough to assume that assessments will have the intended effects just because they have greater face validity. Such assumptions should be checked, and ways of increasing the likelihood of achieving intended outcomes must be identified.

Generalizability and transfer

It is obvious in the case of a multiple-choice test that there is a need to be concerned about the degree to which results generalize to other ways if demonstrating knowledge and understanding. The concerns also apply to performance assessments, however. We need to know the degree to which performance on a written essay, for example, generalizes to performance on other writing tasks. The same can be said for a performance involving a scientific experiment. Do problem-solving skills demonstrated on one lab problem transfer to other problems?

Fairness

Fairness is clearly an issue for any assessment but is a particular concern for high-stakes assessments. Because performance assessments often require substantial amounts of student time, it is usually impractical to administer more than a few problems. Indeed, in some cases, a performance assessment may consist of only a single assignment (e.g., a laboratory experiment). Insensitivity of assessments to differences in background knowledge and the experiences of students outside of school can lead to unfairness. It is critical that such effects be minimized if performance assessments are to be used in making decisions about individual students or schools.

Consequences, generalizability, and fairness are only a few of the criteria to be considered in evaluating performance assessments. Seeking evidence related to these and other criteria will be critical if performance assessments are to realize their promise and not simply come and go as another educational fad. Hard evidence will be needed not only to satisfy skeptics and justify the high cost of performance assessments but also to distinguish between effective and ineffective assessments.

CHAPTER FOUR

Alternatives in Action

The search for more authentic assessments has seen a variety of many promising programs introduced that seem to answer many of the concerns about traditional standardized testing.

Dale Carlson had been director of the highly respected California Assessment Program (CAP) for 17 years when he had a revelation. It came to him during a meeting with teachers: "What we had been doing was very sophisticated but it was also irrelevant. We were missing the whole point."

The teachers told him that when standardized testing time came around, they stopped what they were doing, prepared their classes for the tests, and then went back to teaching again. "Something was very wrong," he said.

It didn't take long for CAP to swing into action. "The easy accountability of the past could no longer be tolerated," a 1990 CAP position paper recalled. "It was essential that we find out what students could actually do." If the outside world demanded that students are taught how to think and understand—not just to recognize more or less correct answers to questions—then there must be a way to test that thinking and understanding.

And there had to be a way to return assessment to its rightful place in the classroom—as a part of instruction.

The new system was based on more than performance assessment. It was teacher involvement every step of the way, resulting in more relevant testing—and teaching — and a rejuvenated teacher corps.

State education officials rallied behind the adage, "What you test is what you get," which was another way of saying that you can't have curriculum reform without testing reform. For those schools that resisted change, a new accountability program would provide the incentive. On the other hand, schools that move into the forefront on their own would be handicapped unless a more appropriate assessment system were adopted.

And thus the stage was set for a statewide "Education Summit" in 1989, which called for CAP to make a major shift to performance-based assessment. Bills were introduced in the legislature, and CAP was off and running.

Performance assessments at work

Here is how the new California performance assessments work in the various subject areas, as reported in a 1990 CAP summary paper titled (probably with pun intended) "California: The State of Assessment."

Writing—

Teachers on the CAP Writing Development Team develop all the testing and instructional materials. For each type of writing assessed, the team develops a special set of writing tasks and a scoring guide that identifies the thinking and writing requirements without being narrow or prescriptive. The essays are scored by several hundred teachers at four regional scoring centers over four to six days.

Special handbooks have been developed to give teachers practical materials for every type of writing assessed, including sample tasks, illustrative essays and related readings.

Does the system work? CAP feels that it has been an "outstanding success." Educators throughout the state "have expressed the belief that no single program has ever had statewide impact on instruction equal to that of the writing assessment." And a study by the Center for the Study of Writing at the University of California (Berkeley) confirmed that.

Most of the teachers surveyed (78 percent) indicated that they assign more writing, and almost all (94 percent) said that they assign a greater variety of writing tasks than before the writing assessment was introduced. Students also report a sizable increase in the number of papers and reports they have written.

These are the types of writing required in the grade 8 assessment, with 45 minutes given for each essay:

- ***Autobiographical incident***—narration of a specific occurrence in a student's life presenting it vividly and stating or implying its significance.
- ***Evaluation***—judgment on the worth of an item, such as a book, movie, or consumer product, supported with reasons and evidence.
- ***Problem solution***—analysis of a specific problem, including proposals and arguments for a solution.
- ***Report of information***—objective presentation of data

The 3 Ps: Performances, Portfolios, and now . . .

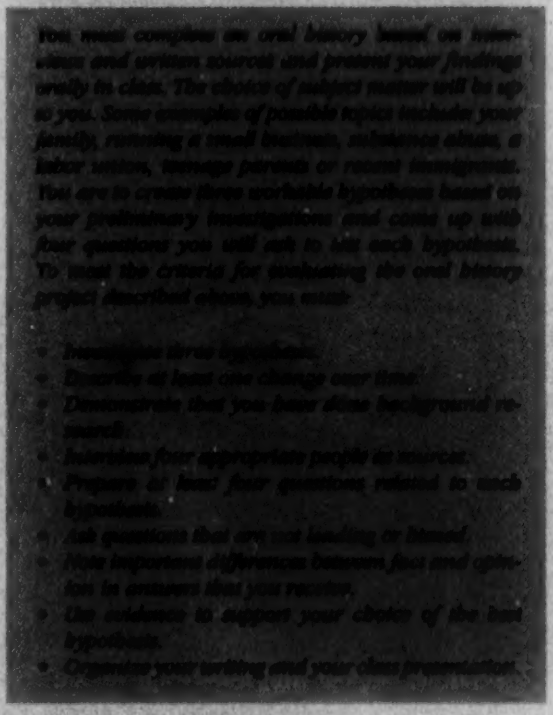
Projects

Projects have long been an integral part of the classroom teacher's arsenal, but they are now being used more frequently and in a wider range of subject areas as performance assessments. Grant Wiggins, one of the nation's leading assessment researchers, offered this advice on turning a classroom project into a performance task.

The project can be to create tangible items in a variety of disciplines from art to woodworking. They can also be presentations. What is important is that the teacher establishes the appropriate steps in their development and encourages many points of self-analysis.

Research projects—individual or cooperative, experiments or oral reports of multiple activities—are becoming more common in primary grades through higher education.

Wiggins offered this example of an oral history project for ninth graders:



collected from observations and research to explain a phenomenon or concept.

- ***Firsthand biography***—presentation of a person the writer knows well, including the significance of that person in the writer's life.
- ***Story***—narration of a fictional story with a plot, characters and scene.
- ***Observational writing***—re-creation of a writer's perceptions of an experience (as an observer rather than participant).
- ***Speculation about causes or effects***—conjecture about the causes or results of an event, trend, or phenomenon.

In grade 12, students are assessed on five of the same writing types as in grade 8—autobiographical incident, evaluation, report of information, observational writing and speculation about causes or effects—but three new ones are substituted for the others:

- ***Interpretation***—interpretation of a text or data supported by reasons and evidence.
- ***Reflective essay***—exploration of a personal reflection and ideas arising from a particular occasion (something observed, a book read, a movie seen).
- ***Controversial issue***—defense of a stand on a controversial issue, supported with sound reasoning and evidence.

The integrated language arts assessments in reading and writing, which were field tested in 1990, delve even further into a student's thinking processes. These new tests ask students to read, write a number of short responses, participate in a small group discussion, and write an essay to prove comprehension.

Mathematics—

California is moving in several new directions in mathematics, in addition to using a limited number of multiple-choice tests.

- ***Free response questions*** at the 12th grade level are complex, and the responses involve more than a single correct answer. Students must make their own assumptions and give reasons for their conclusions. The use of these open-ended questions, moreover, was seen as improving the alignment of assessment with curriculum, because unlike multiple-choice measures, they allow students to demonstrate the depth of their understanding and to solve problems in many different ways.
- ***Mathematical investigations*** are 40- to 60-minute complex problem-solving projects given to individuals or small groups. Students must plan and complete the investigations, answering periodic questions that demonstrate their understanding of mathematical concepts and the thinking that underlies their actions.

- ***Portfolios***, samples of student work collected in the classroom, are designed to evaluate student attainment over a period of time. They include components that are ordinarily difficult to measure using other methods, and may be assessed on a sampling basis statewide.

Results of the free response questions were reported for the first time in 1990 as part of the school, district and statewide reports in mathematics. The 95 high school teachers who were trained to rate them—and scored approximately 24,000 of them in one weekend—were enthu-

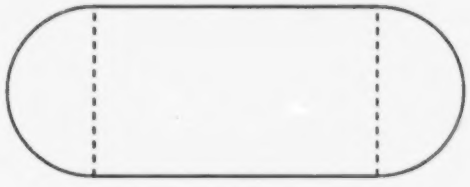
Continued on page 37

Open-Ended Math

Here are some sample open-ended mathematics questions created by the California Assessment Program. The generalized scoring rubric reproduced in this chapter can be used to rate the answers.

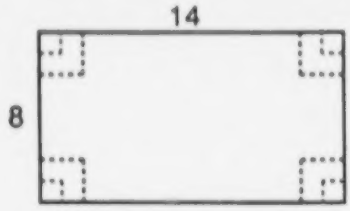
Problem

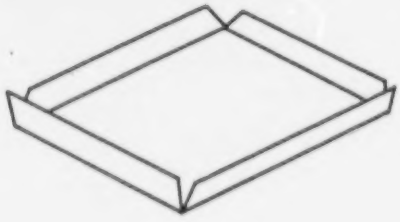
Suppose that a running track is designed with two semicircular ends and two parallel straightaways of 100 yards or more, as shown below.



If the perimeter of the track is to be exactly 440 yards, explain how you would decide how long the straightaways should be and what the resulting diameter of the semicircles would be.

Problem





Henry wants to make a box from an 8 inch by 14 inch piece of cardboard. By cutting the same-sized square from each corner, he can fold up the sides to make a box, as shown.

If he cuts out only squares with whole-number sides, what is the volume of each box he can make?

If he is not restricted to whole-number squares, is there any larger box he can make from the cardboard? Explain.

Problem

Room 15 has 8 tables that measure 3 feet by 6 feet each. The tables are made so that they have to be fastened together in one of three ways:

When all of the eight tables are fastened together, how could they be arranged to give maximum seating? Draw a sketch and explain why you think this arrangement gives the most seating.

Continued from page 35

siastic about their experience. They felt that reading the student responses gave them a unique insight into student thinking, which would enable them to target instruction better.

Science—

The performance tasks field-tested in 1990 had more than 50,000 sixth graders in 1,000 schools engage in hands-on activities in which they were put into a simulated "real life" situation and told to solve a problem. They were asked to observe, classify, sort, detect patterns, infer, formulate hypotheses, and interpret results.

The format consisted of five "stations" with one task per station. Each task took about 10 minutes, and the students rotated through all five stations in one testing session. It was not enough to answer the question. The students had to write a sentence or two on the answer sheet to explain how they knew their answer was right.

History and Social Science—

These tests, scheduled to be developed in 1991, were designed to allow students to demonstrate breadth of learning as well as their ability to clarify issues, recognize relationships, determine causes and effects, interpret evidence, and argue for a position. The idea is not to emphasize facts but to assess a more thorough study of eras, events, and original documents in their social and cultural contexts.

Portfolios in Vermont

Vermont Commissioner of Education Richard P. Mills also credits a group of concerned teachers with inspiring his state's pioneering movement into alternative assessments. They complained at a statewide test proposal hearing that because traditional test scores look acceptable to the public, teachers feel compelled to waste their time—and their students'—to "make the numbers."

Mills reacted quickly. In an article in the December 1989 issue of *The School Administrator*, he wrote, "How confusing it is to coach students in good writing, and then test them by asking them to fill in a multiple-choice test! How perverse to insist they write and rewrite and edit over days, and then test them with a 40-minute writing sample!"

He went to citizen groups and asked for their input on a new approach for measuring student performance. "Could we find an approach that would be a part of teaching and not apart from it? How could we help local educators build sophisticated performance data into current practice?"

The plan earned strong support from Governor Madeline M. Kunin and the state legislature. Vermont launched a statewide system of assessing performance through portfolios of student work, beginning with writing and mathematics, then expanding later into other areas of the curriculum.

The portfolios will be used in conjunction with standardized tests to be developed by Vermont teachers, using Vermont goals, as an "important anchor," in Mills's words. Although these tests had not yet been developed, the plan was to make them performance-based and ensure that they were equivalent tasks administered under the same conditions for each student.

Four writing categories

The writing portfolio required for 1990-91 contains examples of work from the following four categories:

- A poem, play or personal narration.
- A "personal response" to a cultural or sports event, book, math problem, or current issue.
- Prose pieces from classes other than English and language arts.
- A piece that students consider their "best."

Learning beyond facts

The math portfolios contain a range of materials that are designed to demonstrate learning beyond facts and knowledge. In this way, teachers hope to be better able to see examples of the type of higher thinking skills required in real life than they would from a page of addition or subtraction problems.

Student's process

Portfolios also are used to show the process, or evolution, of a piece of work. Just as artists use preliminary sketches, students are encouraged to keep and share notes, outlines, outside sources, and rough drafts that lead to the production of a finished paper.

If the portfolio shows the rough calculations and experimentation that led to the solution of a math problem, teachers can better assess the strengths and needs of the student. Thus they can encourage and reward good thinking—even though it may not have led to the right answer.

More complete

Because of the range of materials in the portfolios, supporters of the program feel that they will provide more complete information about student abilities than the traditional tests.

"This is a much more complex way of looking at performance," said state education planning director W. Ross Brewer. "This shows what are the real things people can do and where they are coming up short."

Continued on page 42

Assessing the Essential Skills

The Arizona Student Assessment Plan (ASAP) is centered on measuring students against broad curricular goals called Essential Skills. Here is how that concept works in a sample performance-based reading assessment developed for the state by the Riverside Publishing Co.:

Grade 8 Reading

Essential Skill (Outcome): Reads and comprehends a poem.

Competency indicators:

- Tells what the poem is about (literal comprehension).
- Makes an inference about the poem (interpretive comprehension).
- Relates the poem to own experience (critical comprehension).
- Gives a personal reaction to the ideas and feelings in the poem (appreciative comprehension).
- Gives a personal reaction to the pattern or imagery in the poem (appreciative comprehension).

Essential Skills (Process):

- States what is already known about content.
- Predicts content.
- Summarizes, clarifies, questions, and predicts while reading.
- Monitors own comprehension and self-corrects when necessary.

The test focuses on the poem "Daffodils" by William Wordsworth, reproduced as follows with three obscure words defined:

Daffodils

I wandered lonely as a cloud
That floats on high o'er vales and hills,
When all at once I saw a crowd,
A host* of golden daffodils;
Beside the lake, beneath the trees,
Fluttering and dancing in the breeze.

Continuous as the stars that shine
And twinkle on the Milky Way,
They stretched in never-ending line
Along the margin of a bay:
Ten thousand saw I at a glance,
Tossing their heads in sprightly dance.

The waves beside them danced; but they
Out-did the sparkling waves in glee:
A poet could not but be gay,
In such a jocund* company:
I gazed—and gazed—but little thought
What wealth the show to me had brought:

For oft, when on my couch I lie
In vacant or in pensive* mood,
They flash upon that inward eye
Which is the bliss of solitude;
And then my heart with pleasure fills,
And dances with the daffodils.

*host—a great number
*jocund—merry
*pensive—thoughtful

In a pre-reading activity, students learn what kind of knowledge they are expected to bring to their reading of a poem. Comprehension questions are then asked about both the poem's content and its formal elements. The students compare daffodils, as described in the poem, to another plant they are familiar with, and finally discuss the poem's meaning.

The test itself begins with a brief paragraph about the Romantic Age and asks a few questions about that. Then the poem is reproduced (with footnotes defining several of the more difficult or obscure words in it). Four exercises follow:

Exercise A:

Complete the following questions.

1. Where do most of the events take place?
2. What is the setting of the last stanza of the poem?
3. What is the mood (feelings evoked) of the poem? (Please circle the letter of the correct response):

   A. trusting B. sad
   C. forgiving D. happy
4. In the third stanza, how do the daffodils compare with the waves?
5. What lasting effects do the daffodils have on the poet?

Exercise B:

The author uses a great deal of imagery (mental pictures) in his poem. For example, he states that there are a great number of golden daffodils fluttering and dancing, and he compares them to the stars in the Milky Way. How do these mental pictures add to your understanding or enjoyment of this poem? Write a paragraph explaining your answer.

Exercise C:

Think about another plant you like very much—a rose, a lily, a cactus, a geranium, a tulip, or any other plant. How does your choice compare with what you've learned about daffodils from reading the poem? Write a paragraph that answers this question. Refer to the Review Check List below to help you write.

Exercise D:

What do you think the author hoped a person would experience in reading a poem? Is there a message in the poem? Keep in mind the selection about the Romantic Age you read on an earlier page. Then write a paragraph that answers those questions. You may look back at the poem or at the reading passage. Use the Review Check List below to help you write.

The Review Check List contains three items:

- ☐ This paragraph clearly explains the three questions asked.
- ☐ This paragraph is written in complete sentences.
- ☐ This paragraph has been checked for correct spelling, capitalization, and punctuation and is written neatly.

Scoring Rubrics:

There are slightly different scoring rubrics for each of the paragraphs required. Following is the one for Exercise D, where students are graded from 0 to 4:

A 4 response

- Answers the questions in a unified manner, using a topic (main idea) sentence.
- Includes supporting ideas.
- Exhibits creative or original thought.
- Interprets the poem in a way which accurately reflects the poem's content, formal qualities, message, and imagery.
- Uses complete sentences.

A 3 response

- Contains a topic (main idea) sentence.
- Is reasonably unified and organized.
- Includes some supporting ideas or examples.
- Includes a literal and appreciative comprehension of message, words, and images which is clear, although the writing may not be fluidly expressed.

A 2 response

- Contains ideas, as opposed to a unified expression of main ideas.
- Organizes in a manner similar to listing.
- Shows weak comprehension of the poem and the author's purpose.
- Exhibits distracting patterns of misspelling, word usage problems, punctuation, and other language and mechanics problems.

A 1 response

- Gives the topic only, providing little, if any, support.
- Provides information which is inaccurate or misrepresentative of the poem and makes little attempt to speculate on the author's purpose.
- Produces sentences which appear random and show a lack of clarity.
- Complicates message with serious language and mechanics problems.

A 0 response

- Indicates that the student has failed to attempt the question.

Sample Rubrics

These are two sample scoring rubrics used by the California Assessment Program in grading one of its writing tests, in Evaluation, and its open-ended mathematics examination. In each case, students are graded as 0 through 6.

Demonstrated Competence

For 6 points: Everything is correct, including the picture (five interior squares), the scale, the sequence of areas, and the rule. The rule may be a formula (equivalent to $1/2^{n-2}$) or recursive, equivalent to $A_1 = 2$, and $A_{n+1} = A_n/2$. Note: The rule "1/2 previous area" without the initial area is *not* adequate.

For 5 points: There is a correct drawing and a correct sequence of areas (e.g., 2, 1, 1/2, 1/4, 1/8) but not a rule.
Or, there is a correct drawing, a correct sequence, and a correct rule except for scale.
Or, there is a modified drawing (not trivial), a consistent sequence to scale, and a consistent rule.

Satisfactory Response

For 4 points: There is a correct drawing for Part a; for Part b, a correct sequence of areas except for the scale derived from mislabeling the original square (e.g., 32, 16, 8, 4, 2); Part c is not attempted.
Or, the problem is modified; but it is nontrivial, and a consistent sequence of areas and a rule (Part c) are given. The scale may still be incorrect.

For 3 points: A correct drawing shows five interior squares; no other correct information is presented.
Or, the problem is altered by a modification of the drawing, producing a trivial problem; a correct solution for Part b is given in accord with the modification.

Inadequate Response

For 2 points: The drawing shows the midpoints of the original square or of subsequent squares, but midpoints are not connected in order. There is no other correct information presented.

For 1 point: Some attempt is made at a solution, but no indication is given of an understanding of midpoints.

Off Track: The student leaves a blank page or writes: "I don't know."

Remark: Misinformation in one part does not reduce points earned in other parts.

Generalized Rubric

Demonstrated Competence

Exemplary Response . . . Rating = 6
Gives a complete response with a clear, coherent, unambiguous, and elegant explanation; includes a clear and simplified diagram; communicates effectively to the identified audience; shows understanding of the open-ended problem's mathematical ideas and processes; identifies all the important elements of the problem; may include examples and counterexamples; presents strong supporting arguments.

Competent Response . . . Rating = 5
Gives a fairly complete response with reasonably clear explanations; may include an appropriate diagram; communicates effectively to the identified audience; shows understanding of the problem's mathematical ideas and processes; identifies the most important elements of the problems; presents solid supporting arguments.

Satisfactory Response

Minor Flaws But Satisfactory . . . Rating = 4
Completes the problem satisfactorily, but the explanation may be muddled; argumentation may be incomplete; diagram may be inappropriate or unclear; understands the underlying mathematical ideas ; uses mathematical ideas effectively.

Serious Flaws But Nearly Satisfactory . . . Rating = 3
Begins the problem appropriately but may fail to complete or may omit significant parts of the problem; may fail to show full understanding of mathematical ideas and processes; may make major computational errors; may misuse or fail to use mathematical terms; response may reflect an inappropriate strategy for solving the problem.

Inadequate Response

Begins, But Fails to Complete Problem . . . Rating = 2
Explanation is not understandable; diagram may be unclear; shows no understanding of the problem situation; may make major computational errors.

Unable to Begin Effectively . . . Rating = 1
Words do not reflect the problem; drawings misrepresent the problem situation; copies parts of the problem but without attempting a solution; fails to indicate which information is appropriate to problem.

No Attempt . . . Rating = 0

Continued from page 37

"My best work"

And the idea of having the students include what they consider their best work in the portfolios in itself encourages students to strive to a higher standard. By contrast, traditional tests encourage a "compliance mentality," according to assessment expert Grant Wiggins, who served as a consultant in Vermont.

Mills concurred about the value of that part of the portfolio. "That's a powerful question—'What's my best piece?'" he said. "In answering it, students have to reveal their own standards." In his peptalk to more than 500 teachers who gathered in October to kick off the program, Wiggins asserted, "Many people will see what you are about to do as the opening shot in a campaign to see if we can have standards without standardization."

The whole curriculum

Arizona put at least one foot into the performance assessment waters in 1990 after a survey verified what many educators knew all along: (1) The standardized test used by the state—the ITBS—covered only about one-quarter of the state's curriculum, and (2) teachers had the tendency to teach primarily what was on the test.

A cooperative effort by the State Department of Education and the legislature resulted in a law that, in effect, required the schools also to evaluate how students achieve on the other three-quarters of the curriculum. Thus teachers would be motivated to teach to the rest of the curriculum as well.

Actually, recalled Lois Easton, director of curriculum and assessment, the law did not start out proposing performance assessments. It was called "Goals for Educational Excellence," and required schools to set high goals and measure how well they were approaching them.

"The trouble was that we already had high goals, but teachers weren't teaching toward them, and we weren't able to measure them because of the pervasiveness of norm-referenced tests," Easton said. "We couldn't do away with those tests altogether, because they do play a role. So we knew we would have to reduce them significantly and then come up with something to take their place. That something *had* to be performance-based assessments."

So as Arizona moved inexorably toward the publication of the first "School Profiles" in 1992, the state's effort differed from the other states pioneering alternative assessments in one major direction: Arizona planned to give each of its 214 local school districts almost total authority and initiative to make the system work. Although the state would be watching over their shoulders, the districts could develop their own programs to help their students meet the state's high curriculum standards known as Essential Skills—as long as they could prove it.

Nevertheless, the State Department of Education got busy right away preparing materials to make sure that everybody moved in the right direction. The first step in the Arizona Student Assessment Program (ASAP) was a pilot program of performance-based assessments administered to a sample of students in reading, writing, and math in grades 3, 8, and 12. These pilot tests were distributed to teachers in the fall of 1990 to help them start planning their instruction and classroom assessments.

Help shape instruction

Starting in the spring of 1991, three versions of pilot tests were sent to the districts to help them shape their instruction for the grades leading up to 3, 8, or 12. The real thing was due to be administered in 1992, and it was not to be a whole lot different from the pilot tests. All the items call for the same processes and activities in new contexts. As an example, a math assessment on the pilot might ask the child to design a playground within a certain lot size. The actual test might call for the design of an amusement park within slightly different dimensions.

"We teach what we value. We test what we value. They should be the same," Easton said.

Local option

Districts had the option of using the pilot tests any way they wish, or not at all. Easton called them "an open-ended pilot." They could be used as local assessment instruments, or as instructional units (which they resemble) or as the basis for a staff development program. They could—and in many cases should—lead to a realignment of curriculum.

The moment of truth will come when 1992 rolls around and the real test, called Form D, is administered. Admittedly, Easton said, the state assessments will only "sample" the state curriculum standards and only at those three grade levels. But the law also requires that the local districts come up with an effective way to prove that all their students can meet all the standards.

In other words, the local districts are being given the time to try out new assessment systems on their own and see how they work. Eventually, they will be required to report on the results of their own assessment systems. As one suggestion, Easton said, they could begin the use of portfolios at the school and district level. "We don't always need to test to find out whether students are making progress in the competencies."

Easton said that in creating the system, everybody was very conscious of the "local control" issue and the state did not want to be heavy-handed. However, starting in 1992, if the test samples show a very different picture of student achievement when compared with the district's own reports, "There may have to be some intervention."

Five-year cycle

The publication of the first School Profiles in 1992 will begin a five-year cycle of accountability.

In addition to the performance assessment component described above, the Profile for each school will include results of a much-reduced administration of norm-referenced testing. These tests—but not all of the batteries—will be given in the fall instead of the spring. "We're doing everything to take away their high-stakes character," Easton said.

The cart and horse revisited

The underlying purpose of all these alternative assessment programs is not so much to create better tests, but to use tests as a way of improving curriculum.

As noted above, the Arizona program was created in response to a legislative mandate for schools to set high goals and measure how well they were approaching them. While most of the early efforts have been in developing tests, Easton mused that what looks like merely a new assessment program may turn out to be an "overhaul of the whole system."

That, Easton feels, may be the real difference between Arizona's approach and everybody else's: "We didn't just sit down and write new tests. We started with the standards that kids ought to be able to meet, and based our tests on them. If that forces schools to revise their curriculum so they teach to our test, we've done our job."

Teaching to the test

There is that phrase again—representing one of the pernicious abuses of traditional norm-referenced tests but now one of the professed goals of alternative assessment.

Once again, Lois Easton: "We want our assessment system to be transparent, not secret and secure. We want to know that students are doing what we want them to do. If the teachers teach to the test, then we'll know the students can read and write and use math for a purpose."

Easton added that "keeping tests secret from the teacher is like asking a band director to prepare for a concert without copies of the sheet music."

Similar echoes could be heard emanating from Maryland, which was busy developing performance tasks as part of its new Maryland School Performance Program (MSPP). The State Board of Education asked Ruth Mitchell, associate director of the Council for Basic Education (CBE), to take a look at the progress of the program.

Her report was glowing: "In comparison to performance objectives in other states, the Maryland program seems to me ingenious in its careful balancing of evaluation and instructional values, courageous in attempting an almost impossible timeline, and powerful in its potential to ensure a high level of curriculum and instruction for all students."

She cited as an example a grade 8 reading/writing/language arts assessment. It consists of a story by Jack London coupled with a short extract from a book on hypothermia, with a series of exercises using each to interpret the other. "It is exemplary," she said. "If this assessment provokes this kind of teaching instead of reading packaged material and answering the questions at the end, it will change curriculum and instruction for the better across the state."

What she didn't like were comments by some curriculum coordinators who urged teachers not to teach to the test. She hoped they would change their tune. "These assessments are worth teaching to," she told the state board, "because students need to learn the high-level content and thinking skills that the assessments demand."

"A child's poor performance—and need for help—should never again be hidden behind a school's high average scores, and a high performing student should not be denied recognition by being obscured among low average scores."

—Maryland Governor's Commission on School Performance

Teacher involvement

What Mitchell did not like in Maryland was that in the rush to complete the program, there was little time to involve classroom teachers so they would become familiar with the assessments before having to administer them. Elsewhere in the nation where the concept has been around for a while, however, teacher involvement is one of the keys to success.

When Carlson is asked whether California's new assessments improve student achievement, he has to be cautious. Without hard data available to prove it, he can only say, "All reports are that it does."

But does the system improve teaching? "Absolutely!" he declared. "It's the best thing that's ever happened to many of our teachers!"

The days they spend in groups developing tests and scoring student papers can be "wonderful professional experiences," Carlson said. "I've had teachers walk up to me afterward and tell me, 'I know so many new things I didn't know before!' Even the best staff development programs haven't been able to get teachers to work together so effectively on a common problem."

Avoiding the "narrowing"

One of the specific concerns about performance assessment is that it can have the effect of narrowing the curriculum because of the time it takes to conduct just one such activity. The California Assessment Program has responded with a procedure designed to maintain the depth of curriculum even though the performance test may be narrow. It is called matrix sampling, in which every pupil at a particular grade level takes a test that is, in fact, a randomly selected subpart of a larger test.

While there would be little value in collecting comparative scores of individual students, the resulting scores for schools, districts and the state as a whole could be of great significance. And since the tests are, in fact, based upon a broad content domain, it would be impossible for teachers to teach to the test by narrowing instruction.

To carry this concept a step further, a state could create a cycle of large scale assessments so that different subject areas would be assessed comprehensively each year. However, the more items that must be created for the larger test, the higher the price tag goes.

Scoring performances...

If one of the alleged advantages of multiple-choice tests is that they can be scored by machine, how can a performance or a portfolio be rated with the same degree of objectivity? The key, say the experts who have done it, is to establish beforehand a detailed scoring guide that should contain all the behaviors that are expected and delineate the degree of skill that must be demonstrated. It may not be enough to tell a rater that a piece of writing must contain elaboration. The amount and nature of the elaboration must be illustrated before there can be fair and consistent rating. And it is just as important that the students themselves know what these guidelines are before they take the test.

Generally, the scoring is done by a cadre of teachers who are first given special training by the assessment specialists. In California and Connecticut, the states that have thus far conducted the most extensive performance testing, the teachers are paid stipends for their services. As noted earlier, to keep its budget down, Michigan used volunteers.

Because each piece of work must be scored by more than one rater, the job can be a mammoth one. Take, for example, California's first writing assessment in 1987. The scoring of 282,155 booklets was completed in eight days in July by 290 readers and leaders.

Open-ended math

The scoring of open-ended mathematics questions raised a particular challenge in California. The committee developing the 1987-88 program grappled with the structure of the items, according to the report issued by CAP in 1989. "A tightly structured question is easily scored but yields little interesting information about mathematical understanding. A broadly structured question allows students to respond creatively, demonstrate the full extent of their mathematical understanding, and display the elegance and originality of their thought processes." The report noted that after struggling with this "dichotomy," they settled on a mixture—two tightly structured questions and three with relatively broad structures.

The committee also debated the method of scoring open-ended math questions. The review noted that two frequently used methods were "analytic" and "holistic." In analytic scoring, the student is graded according to how many specific points are covered in the response. In holistic scoring, judgment is based on the response as a whole rather than on specific features of the answer.

"Holistic scoring focuses on the merits of the response as a whole, rather than on how well the writing conforms to predetermined analytic schemes," the report said. "Scoring holistically leaves the evaluator free, as we hope a teacher would be, to respond to the student's entire thinking process rather than concentrate on the number of points to be marked off according to a scheme." A general scoring rubric for open-ended math appears later in this chapter.

"Range-finder" technique

Joan Boykoff Baron, who has directed the Connecticut Mastery Test Program since its inception in 1987, said that her state's rating system, based upon the successful "range-finder" technique used by many states in their writing assessment programs, is highly accurate in other subjects as well.

"You have to train teachers to internalize samples of work at the various score points (in Connecticut it's 1, 2, 3 and 4). With an hour or so of training, you can get really strong agreement among the trainers." Each paper is graded by two raters, who must demonstrate their skill before they

can participate. "If any of the raters fail to meet the criteria of accuracy that we previously established, they will not be permitted to score the exams," she said.

Those who do the training have to make sure that although the assignment, and the raters, may change from year to year, the standard doesn't. Baron said that's built into the system too, so in the annual report, progress or lack of it is shown by changes in the distribution of scores.

. . . and portfolios

Vermont will train teachers to serve as assessors. Each year they will evaluate a sample of portfolios from around the state against at least seven criteria to prepare a detailed analysis of student progress.

In writing, for example, they will be looking at the degree to which the organization suits the writer's purpose, how the use of detail adds to clarity, as well as progress over time. In math, the materials will have to demonstrate the student's ability to learn and understand beyond the level of facts and knowledge. All students will receive individual "portfolio checks" in one-on-one consultations with their teachers.

Results from the portfolio and "best piece" assessments will be reported in what Mills calls a "school report day," a kind of town meeting where parents will be able to see their children's work and its evaluations. This means that people will be discussing real examples of real progress instead of talking theoretically about fuzzy concepts like a 550 on the Scholastic Aptitude Test or 54th percentile on a norm-referenced test.

With the portfolios as the keystone of Vermont's reporting program, "People can not only hear about the results," Mills said, "they can see and touch the work students are doing."

Rather than reduce the portfolios to numbers, teachers can use the work in them as teaching tools, asking students to go back to an earlier writing and rework it based on newly developed abilities. And of course the contents of a portfolio can become the occasion for an impromptu discussion. Surely they don't want to reduce the richness of a whole year's work to a single score.

The answer has to be "professional judgment," Mills said. "Judgment has to be calibrated against standards the profession shares and the public understands and supports."

The bottom line

The increased cost of performance assessment is undeniable. Instead of buying a standardized package from a publisher and administering it with little ado, a school district has to commit people and time to developing activities and assessments and training staff in their use. Then comes the bill for training evaluators and conducting the scoring activities.

No less a supporter of performance assessment than Dale Carlson of California conceded that it is a lot more expensive. His writing assessment, he estimated, costs $5 a student compared with $1 the old way. But $3 of the $5 goes to teachers for reading the papers, Carlson pointed out.

"I prefer to call that staff development money," he said. "Teacher involvement brings you support, new ideas, and most of all rejuvenation. It's money well spent!" Although the costs in other subject areas may be dramatically higher, Carlson still feels that every cent spent on the new assessments is improving education. "Every cent we spent on standardized testing was a total waste!"

"Atmosphere of trust"

Baron has no doubts about the value of Connecticut's program—to the individual student and school all the way to the commissioner of education and even beyond. "We've built up an atmosphere of trust over time," she said. And while she won't claim any direct correlation, she pointed out that the commissioner reports the data and "the legislature has been supportive of education in our state."

And this is not just off the top of her head. In a 1990 paper Baron wrote for the American Association for the Advancement of Science, she reported that researchers studying the Connecticut experience concluded that policy makers sometimes place higher value on performance data than on multiple-choice tests when both are available.

"Such results indicate that policy makers are ready to pay attention to the results of performance assessments and act on them," Baron wrote. She exhorted her colleagues, "It is time to inform public policy better by developing a wide array of performance assessments that will provide valid and reliable information not only about what students know in limited multiple-choice contexts, but also about what they can do in more realistic settings."

The role of technology

One of the few fields more dynamic than educational assessment is educational technology, and it is probably here where many solutions will eventually be found. While the first picture that the word "technology" conjures up is that of a computer, television is also becoming a more effective instructional tool. A fifth-grade magnet school in Indianapolis, Indiana, illustrates its potential as an assessment tool.

The Key School, based upon the concept of "multiple intelligences" of Harvard's Howard Gardner, uses as its evaluation instrument television portfolios containing individual student interviews, oral exams, and projects. These videos are kept on file to be viewed by students, teachers, and parents alike.

Through the decade of the 80s, the computer became almost commonplace in many classrooms. Commercially available software enables the teacher to tailor test items to individual students and to determine each student's performance level almost instantaneously. Some of these programs automatically redirect instruction to the student's demonstrated area of need. The instructional advantage here is that the student and the teacher get instantaneous feedback, and help can be on its way in seconds.

A similar scenario is played out on a districtwide basis where the school system contracts with a commercial developer of a computer-driven instructional program. As the students work out problems on their terminals, an individual data base is created and updated for teachers and principals to use in planning further instruction. Those who invest in these programs consider them highly successful in motivating and teaching their students. There is, however, a higher cost associated with them, and the district has to be willing to relinquish many curricular decisions to the developers. In time, however, these programs have become more adaptable to local needs.

Computer-adaptive testing

While technology, including the concept of interactive video, is being used with increasing success in local assessment efforts, computers are also finding a niche in large-scale standardized testing settings as well. In the typical pencil-and-paper standardized test, a student is confronted by a range of questions including those that are too easy and too difficult for him or her. On the other hand, computer-adaptive software continually evaluates the student's answers and automatically moves to tasks at a higher or lower level of difficulty depending on how the student performs. Both the level of difficulty and the number correct are considered in the scoring, which the computer completes in milliseconds. Depending on the network configuration, reports are available by item, by student, and by group.

Computer technology is also capable of solving many of the potential problems of cheating or teaching to the test caused by the use of the same questions year after year.

Before too long, computers will be randomly selecting test items from a huge data base and asking students both multiple-choice and open-ended questions. Scoring will be accomplished far more efficiently and accurately than the scanners that read the bubbles.

To many school districts, these eventualities may seem light years away. This fact hit home on one of the AASA survey forms where it asked for the respondent's name, address, phone number, and fax number. One superintendent filled in the fax number space with big block letters: "Are you kidding??!!!" For that district, and no doubt many others, technology—especially for assessment—is a high-cost item that will not be affordable for a long time to come.

States to lead

While it is unlikely that individual school districts will be able to launch major new initiatives of their own during the belt-tightening of the early 90s, state legislatures are likely to keep the movement going strong. Maryland's Governor William Donald Schaefer, in taking his recommendations to the General Assembly in 1991, argued that his state was already spending a fortune on education—but without a clear idea of whether it was doing any good or not. The Maryland School Performance Program was one of very few new programs that survived the legislature's budget axe.

"Maybe we'll look at the remaining standardized tests and decide to do away with them and keep the new program going. I have serious doubts whether anybody really uses them anyway. The teachers say, 'I only do it because the district insists.' The district says, 'We only do it for the teachers.' What do these tests accomplish anyway?"

—Dale Carlson

A Sample Performance Assessment

The Soda Task—Is it sugared or diet soda?

Students are given two samples of soda, one sugared and the other diet. Their assignment: figure out which one is which. Sorry: No tasting!

This is a chemistry performance task pilot-tested by the state of Connecticut in the 1990-91 school year. It was one of many tasks in science and mathematics developed by committees of teachers as part of that state's pioneering effort toward alternative assessment.

The teachers who rate a performance task must have in front of them clear guidelines that detail all the behaviors to be expected of the students and the degree of skill to be demonstrated. Finally, there must be a systematic way of documenting those skills as they are demonstrated.

To illustrate all the essential factors that made the diet soda assignment a true performance assessment, a team of teachers came up with the following script:

THE SODA TASK

Summary of the task: Students are asked to identify samples of soda as being diet or regular based on their physical and/or chemical properties.

Course: Chemistry

Grade/Level: 11-12/Medium

Curriculum topic: Physical and Chemical Properties.
Identification of matter.

Prerequisite knowledge: Students should have some background knowledge on physical and chemical properties and identification of matter.

Equipment needed: Regular soda, diet soda, wire gauze, heat source, graduated cylinders, beakers, tripods, safety glasses, aprons.

Optional: Yeast, Benedict's solution, and sugar dipstick tests.

References: Merck Index, CRC Handbook of Chemistry and Physics, chemical dictionary, and chemistry textbooks.

Suggested length of time: 3-4 class periods.

Participation: Small group and individual.

PART I:

Getting Started by Yourself

You will be given two samples of soda; one regular soda containing sugar and the diet soda containing an artificial sweetener. Your task is to identify each sample as diet or regular. You must base this decision on the physical and/or chemical properties of the two different types of soda. *As in any chemistry experiment, you are not allowed to taste any of the samples.*

Come up with a list of at least three possible ways to identify between the sodas. Explain why you chose them.

PART II:

Group Work

1. Brainstorm a list of possible tests that might help your group differentiate between the two types of soda. Choose two of these tests to carry out.
2. Design an experimental plan for these two tests.
3. Submit your design to your teacher for approval.
4. After getting approval from your teacher, carry out your experiments.
5. Prepare a report of the results of your experiment.
6. Make an oral presentation summarizing the results of your experiments for the class.

PART III:

Finishing by Yourself

If you were given two samples of water, one of which is salt water and the other fresh water, which tests that your class tried out for the sodas would be useful in differentiating between the two? Which tests might not be useful? Can you think of any tests that your group did not do that might be appropriate for this problem? Explain your answers fully.

Notes to the Teacher

Safety considerations: Normal laboratory safety procedures should be followed. Students should wear safety goggles and aprons.

Objectives: Students should be able to:

- Identify and apply physical and/or chemical properties for the purpose of identification.
- Use and make measurements using appropriate units.
- Formulate predictions based on prior knowledge.
- Identify information and steps needed to solve a problem.
- Test predictions.
- Gather data pertinent to a problem.
- Make references based on pertinent data.
- Draw reasonable conclusions and defend them rationally.
- Communicate the strategies and out-

comes of a study through written means.
- Orally communicate the strategies and outcomes of a study.
- Work cooperatively in a group.

Prior preparation. The teacher will have to make up the unknown samples of soda before the actual experiment. These samples should be numbered and recorded so the teacher can verify student results. The teacher will also need to have other necessary equipment available.

Information needed: Some type of clear soda (7-Up, etc.) should be used to obtain the best results. Students should be given an ample supply of soda to complete their tests.

If students choose to use the urine test using glucose dipsticks, they need to show, using references, that the test is actually not for glucose, but for a reducing sugar. Fructose is typically used in regular soda.

Students may not be able to identify differences in the two samples based on density, freezing point, or boiling point since the differences are very slight. (The regular soda should have a greater density, higher boiling point, and lower freezing point.) However, students should not be discouraged from suggesting and trying these tests since they could be valid properties for identifying substances.

The use of the "sticky test" will provide students with reliable results since regular soda contains sugar, while diet soda does not.

Adding yeast to the two sodas will also show a difference. The yeast will react with the sugared soda to give off a distinct odor as well as gas bubbles. Gentle warming will accentuate this.

The benedict's solution test will be positive for regular soda, producing a reddish precipitate with slight warming.

Urine glucose test strips will be positive for regular soda since fructose is also a reducing sugar.

Sulfuric acid will react with the sugars to produce a caramel.

Adding salt to the diet soda causes more fizzing than the regular soda.

Conductivity of the two sodas may be explored.

Sugar has a greater solubility in diet soda than in regular soda.

If the two sodas are partially evaporated, the regular soda will leave more residue. (If students evaporate the soda completely, a black residue will be left that may be difficult to remove from glassware.)

Diet soda may have a stronger aroma than regular soda.

Students may observe that the two sodas differ in color or amount of fizz. Diet soda may have more and larger bubbles.

Techniques: The teacher may want to put out various equipment or materials that are not necessarily needed for the experiment as detractors for the students. The students may find innovative ways to use these materials that were not thought of previously.

The teacher should display samples of the sodas at the beginning of the task, but should not have any laboratory materials visible until after the students have reached

Objectives Rating Form—Group

The group form is a grid *for each student* on which the teacher checks Excellent, Good, or Needs Improvement for each of the 10 objectives. There are also boxes next to each objective to indicate where to find evidence for each objective: the group report, oral presentation, teacher observation, or other.

The objective column begins, "The group should be able to . . ." and lists the 10 objectives specified above in the notes to the teacher.

1. identify and apply physical and/or chemical properties for the purpose of identification
2. formulate predictions based on prior knowledge
3. identify information and steps needed to solve a problem
4. test predictions
5. gather data pertinent to a problem
6. make inferences based on pertinent data
7. draw reasonable conclusions and defend them rationally
8. communicate the strategies and outcomes of a study through written means
9. orally communicate the strategies and outcomes of a study
10. collaborated effectively.

Objectives Rating Form—Individual

This grid takes three of the objectives from the group chart (Nos. 1, 3, and 8) and asks the teacher to check E, G, or N.I., and fill in where to find the evidence. The objectives complete the sentence, "The student should be able to . . ."

CHAPTER FIVE

Beware the Bandwagon

While the burgeoning field of authentic assessment claims many converts among its ranks, there remain many skeptics who warn against getting carried away. The answer, they suggest, may lie in a marriage between the two camps.

Chester F. Finn, professor at Vanderbilt University and former assistant secretary of education for the Office of Educational Research and Improvement, has been credited with a quote that has become a classic. Performance assessment, he said, is "like Star Wars: The idea remains to be demonstrated as feasible."

Just as the movement seemed to have reached runaway proportions, a cautionary word was sounded in early 1991 at a major conference of assessment specialists in Los Angeles: Before jumping on the bandwagon, stop and think. Nobody has yet found a cost-effective way to produce a performance assessment that is reliable for accountability purposes. In fact, the argument went, scoring continues to be overwhelmingly subjective and is therefore unreliable for such purposes as denying promotions or diplomas.

Commenting on his survey of school districts and agencies about their authentic assessments, Walter D. Hathaway, director of research and evaluation for the Portland, Oregon, Public Schools, reported some "promising" steps being taken. But he suggested that "perhaps it is time to slow down and take stock of where the authentic assessment movement is leading and assess the extent to which public education will be likely to benefit from this rapidly growing trend."

Noting that "standardized test bashing has become so fashionable" that it is becoming hard to find anybody in education who wants to defend them, Hathaway said that the alternatives are not at all clear. "Large-scale applications of more 'authentic' approaches to educational assessment . . . have not been proven to be as valid, reliable, objective, accurate, or useful for system-wide assessment as the standardized measures they replace, even though they are many times more costly."

"Skeptics, not opponents"

While the word "opponent" may be proper in referring to the anti-norm-referenced tests movement, those who speak out against the rush of performance assessments would rather be considered skeptics. "I'm not anti-performance assessment," said Edward A. Mehrens, professor of educational measurement at Michigan State University, "I'm just anti-bandwagon." He contended that the authentic assessment proponents have yet to prove their case that they can do better than multiple-choice testing, and until they do, he said, he is willing to use the word "fad" to describe the movement.

Similarly, Joy A. Frechtling, director of educational accountability for the Montgomery County, Maryland, Public Schools, who is active in national measurement circles, likened the rush toward authentic assessment to a crusade for the Holy Grail. "People are doing a 180 now without a solid foundation," she said, "especially in high-stakes decision-making situations. It really worries me, because the rhetoric is so far ahead of the factual information." Her favorite word for the leaders of the movement appears in the title of one of her papers: "Moonstruck."

The problems

What are the problems these "skeptics" see? In addition to the most common concern involving the cost of developing, administering and scoring these assessments, here are some of the other disadvantages cited by people like Hathaway, Mehrens and Frechtling:

- Performance assessments are very time-consuming; they may extend over days or even weeks.
- They tend to provide a multi-dimensional look at a particular skill or area; thus breadth of coverage may have to be sacrificed for depth.
- The development of alternative assessments requires special skills and sophistication on the part of instructors and raters, far beyond what is currently expected of today's classroom teachers.
- Scoring of performances or portfolios is complex, and looks at process as well as product. The "score" may actually be a series of scores or a narrative. For this reason

as well as the inherent subjectivity in scoring, systemwide assessment cannot be validly compared over time. Yet that is exactly what policy makers need to do so they can modify programs that are not working well.

- Performance assessments often include the classroom teacher as a critical factor in the scoring process. This could introduce subjectivity into the scoring process.
- Given the time available for the exercises, only a small number of items can be tapped in a performance "test." This could lead to a "narrowing of curriculum" far surpassing the effects of the paper-and-pencil multiple-choice exam they are supposed to replace. California's matrix sampling solution requires a considerable amount of money to develop many times the number of performance tasks than actually needed at one time.
- Because of the limited number of items but the long time it takes to develop them, there is the potential problem of test security. Teachers can easily remember what will be on the test and be tempted to teach to it.
- There has been no proof that authentic assessments correct the oft-cited concerns of bias and teaching to the test leveled against standardized tests. Mehrens in particular has argued strongly that these abuses can happen just as readily with performance tests and have more impact in inflating scores because of the limited number of items on the test.

"We teach what we value. We test what we value. They should be the same."

—Lois Easton

Some teachers skeptical

Even a long-time proponent like Connecticut's Joan Baron conceded that exporting a performance assessment program is not going to be a piece of cake. Even though the pilot program Baron directed (sponsored by the National Science Foundation in 1989) went smoothly, and many of the cooperating teachers felt it showed great promise for students, they suspected that many of their colleagues would be much less receptive to such innovations, primarily because of the time pressures involved.

Whatever value such a system has, Baron wrote, "The participating project teachers felt that the time required to develop, prepare, carry out, and evaluate a performance task was formidable."

When Connecticut science supervisors were queried about the possibilities at a meeting in the spring of 1990, they estimated that only 15 percent of their 258 science teachers would be *very receptive* to using such measures. They said about 50 percent would be *somewhat receptive* and 30 percent *not at all receptive*.

This didn't surprise Baron, but she is confident that in the long run, performance-based tests will play a major role in an expanded repertoire of teachers' assessment practices and state assessment programs.

The pendulum swings

The metaphor describing educational assessment is more a "pendulum" to Lew Pike, test development specialist for the Fairfax County, Virginia, Public Schools. He decried the "either-or mentality" of many of his colleagues and hopes that his school district will be able to "occupy the middle ground before it's obliterated."

But as the winds of change swirl through the education establishment, observers with an historical perspective fear that the pendulum may have already swung full circle.

For example, Professor Shepard pointed out, it was precisely with procedures like California's landmark system of matrix sampling—such as testing 10 subject areas in four-year cycles—that the National Assessment of Educational Progress started 20 years ago.

It didn't take long, however, for political realities to set in. Cost constraints —and the demand for accountability—gradually forced a shift to more frequent assessments in the basics. The result: a single score for each subject area to produce a national report card.

Walter Haney and George Madaus of Boston College, attempting to put assessment movements into a longer perspective, wrote that "standardized tests became popular in part because of problems with these older forms of evaluation." What were "these older forms of evaluation?" The phrase had not been coined yet, but weren't they "performance assessments?"

One of the advantages that precipitated the proliferation of standardized testing was its built-in yardstick that enabled objective comparisons of students from different teachers, schools, states, and countries, they added. Underline that word "objective," at least as it relates to scoring. If human beings, albeit well-trained ones, are ultimately responsible for judging a child's proficiency, there will always be more room for criticism.

Madaus sees the problems of validity and reliability as solvable, but not until more states and more districts put alternatives into place. "You don't solve the problems until you get involved in implementation and get data you can work with," he said. "We don't have a big enough sample base to work out problems." Without such evidence, he fears, performance assessment could "become a fad like creativity was 10 or 15 years ago or computer-based instruction was a decade ago."

Even worse, says H.D. Hoover, director of the Iowa Basic Skills Testing Program, is the fear that the pendulum will swing back and schools will place even more emphasis than they do now on traditional tests. "I worry that if we over-rely on something, we'll jump in the other direction,"

Hoover said. "In the testing movement, if in fact performance assessment in most cases is promising more than it can deliver, the advocates are going to kill the very thing they are for."

Publishers fill the breach

Although the phrase has been tossed around frequently in the heat of the debate, few educators feel that authentic assessment will become "just another fad."

America's educators are too resourceful—and America's businesspeople too tenacious—to let that happen. There is, indeed, someone for school districts to turn to for a resolution—the same people who got them into this situation in the first place.

The commercial publishers of the norm-referenced standardized tests have been feeling the pressure for some time. And while they continue to record healthy sales of their tried-and-true instruments, they have also decided to go with the flow.

Virtually all of them have put their considerable expertise and resources to work to create systems of performance assessments and open-ended questions and to develop tests that integrate these with multiple-choice questions. Rather than fight the trend, they are encouraging it. CTB/McGraw-Hill, for example, is working under contract with Maryland to help develop that state's new performance assessment tasks; and Riverside Publishing is helping Arizona develop its new assessment forms.

Salvation of the movement

Speaking from a state that did it all on its own, Dale Carlson of the California Assessment Program sees this as the salvation of the movement: "The test publishers are not leading the charge, but they are coming around," Carlson said. The way he sees it, the test publishers can do the technical work in writing the tests, and the local school systems can develop the scoring instruments and train their teachers on how to use them.

In that way, he added, the new assessments can still be purchased from the same publishers for far less money than it would cost to develop them in-house—perhaps not much more than they are paying for the norm-referenced tests. The added cost, of course, would be for scoring, but Carlson insists that the extra salary given to those teachers "is staff development money more than testing money," and will pay for itself many times over. He feels one of the greatest benefits California has realized in its performance assessment program is the "rejuvenation" of the teachers involved in the group scoring activities.

A common curriculum?

Some people might argue that a nationally standardized test—whether norm-referenced or performance-based—cannot truly reflect each local curriculum, but Carlson disagrees.

"There isn't that much variation," he said. "There is some provincialism; but on the whole, there is a common curriculum. Everybody's desired outcomes are pretty much the same, thanks to the influence of the national teacher organizations."

The question is if the scores resulting from such a system can be used the same way as norm-referenced tests to provide yardsticks from district-to-district. It works in California because the state itself controls the tests and trains and supervises the raters. Whether the standardization will permit a valid comparison between independent jurisdictions is another question.

The best of both worlds

In the final analysis, many experts in the field of assessment feel there are advantages both to norm- and criterion-referenced standardized testing and the new breed of authentic assessments. One kind does some things better than the other. The key, said Pike, is to set the curriculum objectives first and then use the assessment best suited to meet that objective.

Frechtling cautioned that "we may not want to be so quick to reject multiple-choice assessments and other paper-and-pencil measures that have made up traditional testing batteries. Indeed, to try to provide a balance between breadth and depth, subjectivity and objectivity, and complex vs. reductionistic reporting schemes, we may want to combine what might be considered the best aspect of multiple-choice testing and performance assessments. There is a good argument to be made for testing different types of knowledge in different ways and providing for what might be considered audits of knowledge, as well as case studies of skills."

And Hathaway said the same sort of thing in somewhat more poetic language in a paper prepared for delivery at a major conference in 1991: "The safe route between the Scylla of traditional 'stifling,' 'limiting,' 'distorting,' objective multiple-choice standardized testing and the Charybdis of alternative, 'fuzzy,' 'ungeneralizable,' 'unaggregatable' subjective methods of assessment is to find ways to adapt and get more useful information from the former while developing more useful and cost-effective versions of the latter."

Hathaway feels that, if anything, the controversy has had "more curriculum people talking to more evaluation people, and that has to be good." As far as a solution is concerned,

he proposed an "interim compromise" in a 1991 paper co-authored by Jacob Hacker, also of Portland, that would do the following:

1. Accept, promote and practice the belief that multiple measures are better than single ones, especially in measuring gain from one assessment to another versus one-time levels of performance.
2. Embrace the desirability of more authentic measures whenever possible and cost-effective.
3. Encourage and support development and use of more authentic assessment techniques by teachers in their classrooms for assessing and monitoring their students' progress and their needs for further learning opportunities and experience.
4. Work with curriculum and instructional professionals to modify standardized, multiple-choice testing systems so that they:
   - Develop and include instruments which assess new dimensions such as context and prior knowledge as the curriculum being assessed suggests.
   - Develop and add to standardized assessment systems more "authentic" items, e.g., for reading use long passages of connected, meaningful text from "published" materials and ask multiple questions, at least some of which tap higher-order thinking skills; or for mathematics, use everyday problems and permit use of calculators for all items except those that assess computation and estimation.
   - Add open-ended, extended, non-multiple-choice items to standardized multiple-choice tests. For example, on a mathematics test pose problems and give students time and space to work out the answers and a place on the answer sheet to code in their responses. Use information from such adaptations in an integrated fashion along with the traditional scores.
   - Add additional authentic "items" in systemwide assessments using a matrix sampling design to lend depth of insight into the meaning of large group, aggregate data while maintaining cost effectiveness.

In his summary, Hathaway cautioned that "the controversy over standardized vs. "authentic" assessments unfortunately diverts attention from the more important mission. Tests and assessment are only tools; they can be either valuable or worthless depending on where, when and how they are used." The solution to the dilemma is not as simple as saying "no more standardized tests," he said. "Perhaps we should be saying, 'No more closed doors' and 'no more closed minds.'"

"Which, if any, of the new classroom-focused and developed methods can produce cost-effective information that will be useful not only for the instruction of individual students, but which can also be made sufficiently accurate, objective, and discriminative as to be meaningfully aggregated to validly represent degrees and differences in student group performance and in program effectiveness at the school, state, and national levels?"

—Walter D. Hathaway and Jacob Hacker

An interview with Greg Anrig

Testing: Neither the Problem nor the Solution

As president of the Educational Testing Service (ETS), Gregory R. Anrig concedes that it may sound odd for him to be warning people not to test so much. But that was the gist of his testimony before a Congressional committee in 1991 on the issue of a proposed national test. To Anrig, that concept was illustrative of the three bandwagons that were picking up speed in the early 1990s:

1. The idea that one test can serve both the needs of accountability and instruction. ("Can't be done," he said.)
2. The belief that all testing in the past was bad and everything in the future will be good. ("Neither statement is true.")
3. The belief that the way to reform education is through testing. ("Assessment is secondary; first you have to decide what it is you want people to do.")

The idea that more testing will improve achievement has been disproved over the past 20 years, he says. By the end of the 1970s, 38 states had some kind of mandated testing program; by the end of the 80s the total was up to 47. "But student performance was just as inadequate."

Movements don't work either. In the 1970s, the cry was "Back to the basics!" The basics are indeed back, he said, but the level hasn't increased much since then. The 80s were the decade of minimum competency, and that was the problem: the competency stayed at the minimum level—"not good enough for anything."

The decade of standards

When it comes to the 90s, however, Anrig is much more confident. He sees this as the decade of "world class standards—reaching for excellence." The problem is how to get there from here.

Rather than emphasize tests, he would rather emphasize standards. "We have to agree what we expect of students and teachers—not a fixed point, but in the context of standards and levels of attainment."

These standards, he stressed, should be "nationwide," but not "national." They should not be set by any government, but they can be established by national teacher associations in consultation with education experts and political and business leaders, as recently done by the National Council of Teachers of Mathematics.

These standards, then, should be implemented on the local level by local educators, who would then develop ways to measure progress. "It should be an educational issue, meaning that the results are used to improve instruction," he warned, "not a subject for accountability."

What can be done about the demands —justifiable demands, Anrig pointed out —for accountability? "We must find suitable methods that are not as costly as those in place today," he said. He defined "costly" in two ways: (1) in money and (2) in time that could be more effectively used for instruction.

"We have to agree what we expect of students and teachers—not a fixed point, but in the context of standards and levels of attainment."

Accountability testing, using more traditional methods, can be implemented at the state level, he said, as long as it is consistent with what local schools are doing. Performance assessment is ideal for instructional testing at the local level.

"It's fashionable today to condemn standardized multiple- choice-tests because students are weak in higher order thinking skills," he pointed out. What these people don't recognize is that it was the standardized multiple-choice National Assessment of Educational Progress (NAEP) administered by ETS that "uncovered that problem in the first place." Standardized tests still provide useful information, he emphasized, when properly used. Multiple-choice tests can indeed tax students' thinking. And they are certainly more economical.

Sampling works

It may be, Anrig said, that traditional assessment methods like NAEP, using small samples of the student population, will be sufficient to meet the needs for accountability. "If the national media can project with confidence the election of the president of the United States by polling 1,500 voters, who says we can't learn a lot about education by testing 100,000 students?" he asked.

He welcomes the addition of performance assessments into the testing arena, noting that half of new NAEP assessments will be performance-based. In addition to their instructional value, he said, this technique may some day be useful for large-scale accountability purposes as technological advances make possible new ways of measuring increasingly complex performance skills.

For the moment, however, the big question is how to improve student progress. The answer, as he has said many times, is not to give new tests, especially those in which the blame falls on the students, as with today's accountability testing.

"First decide what the students should know," Anrig said, "and make sure they have had a chance to learn it. Then test to make sure they did." Anrig's message is that if they didn't learn, that is more properly an instructional rather than a political issue.

CHAPTER SIX

Assessment Begins at Home

Before jumping on any of the new bandwagons, districts should ask themselves whether they have a comprehensive assessment program, or just give tests.

The allure of new assessment techniques has the potential to create a testing hodge-podge at the local district level. Many school systems, reacting to federal or state mandates as well as local concerns, have been adding tests here and there without first trying to determine where there is redundancy and where there is irrelevancy.

When critics cry that today's students are being grossly overtested, they are talking about the tendency over the last two decades to pile test upon test and still not show any noticeable improvement in student achievement. It will only get worse if a school district adopts an appealing new alternative before analyzing what it is already doing—and why. Whatever the district does should be part of a *comprehensive* program.

Developing a comprehensive program

Whether the challenge is to develop a new program or revise an old one, district administrators should be prepared to make two overriding commitments:

- The district's instructional goals and instructional objectives must remain paramount. Assessments should be designed to measure achievement against those goals, not against any arbitrary standard adopted for convenience.
- All segments of the school community must be involved in the process at one stage or another. This means administrative staff, teachers, parents, and other citizens. Full acceptance calls for a feeling of "ownership" by everybody involved.

Here is a suggested ten-step plan for implementing a comprehensive assessment program at the local district level. Note the references to an "assessment specialist." If the district does not have such a trained and experienced person on its payroll, it should consider hiring one or, at the very least, contracting with a skilled consultant for the early steps. In some cases, the district may be able to work with a testing expert from a regional service agency.

1. Assess the needs

Curriculum and assessment specialists should confer first with teachers and principals to determine the *real* assessment needs of the school district. They should be able to articulate what kind of data they want and what they intend to do with it. The information needs of the superintendent and board of education should be added to the mix. The combined needs will probably run the gamut: Norm-referenced data for policy makers, criterion-referenced data for remediation, and performance-assessment data for direct classroom instruction. There can be a place for each in a school district.

2. Develop the instruments

Once those needs are compiled, the assessment and instructional specialists should work together to determine the best instruments or methods to achieve them. Those tests mandated by the state or other agencies should be analyzed to see which of the needs they can fulfill. The rest of the existing testing program should similarly be scrutinized to identify instruments to be retained or replaced. Then a decision must be made whether commercial tests will be adequate or if local tests should be developed. For example, there is no point in launching an expensive developmental effort if the school board wants an idea of how local students compare with others elsewhere. A sampling of students on a commercial norm-referenced test would do that well enough.

3. Review all instruments

Specialists should review all tests being considered for purchase to make sure they address the school district's curriculum goals and philosophy. For those assessments that are to be developed, a plan for test specifications

should be developed and reviewed. For example, if a district wants to develop criterion-referenced tests in science, a committee of science educators and assessment specialists should determine the content to be covered, the method of assessment (multiple-choice, demonstration, etc.), the scoring criteria, and the reporting system.

4. Train test developers

For tests that are to be developed locally, once the specifications have been written, the teachers and other involved staff members must be trained to write test items to match the specifications. A summer workshop can easily produce more than the number of items necessary for a test. The items must then be reviewed for technical quality by the assessment specialists, for content validity by instructional specialists, and for possible bias (see below). After all items have been edited to address the concerns of all the reviewers, the test forms can be constructed.

5. Inform all the staff

While the tests are being selected and/or developed, the administrative and supervisory staff must be kept informed about the reasons for the various assessments, the steps being taken, and the potential impact on the classroom. The administrators should be responsible for communicating the same information to their staffs. Nobody who works for the district should be uninformed about the program—especially those staff members who will be required to implement it or report the results. Every effort should be made to detect any skepticism or fears on the part of staff members and to address those concerns by explaining all the factors that went into the decision-making process. The public appreciates a united front when it comes to discussing any school procedures.

"For years, students of the senior class were required to read 'Phileopolis' and answer questions about its meaning, etc. Teachers were not required to do so, but simply marked according to the correct answers supplied by Miss Quist, including (1) To extend the benefits of civilization and religion to all peoples, (2) No, (3) Plato, and (4) A wilderness cannot satisfy the hunger for beauty and learning, once awakened. The test was the same from year to year, and once the seniors found the answers and passed them to the juniors, nobody read 'Phileopolis' anymore."

—Lake Wobegon Days *by Garrison Keillor*

6. Begin specific training

This is the time for specific training to begin for designated staff members in the administration of the tests. School-based administrators should be conversant with the various aspects of test construction and administration, but it is advisable to have at least one person in each building act as a *school test coordinator.* This person should be trained to handle all test materials, instruct teachers in the actual administration of the tests, monitor the quality of the administration during testing, and process and return all materials so that scores will be tallied in the most efficient manner possible.

7. Identify training needs

Now that the assessments have been determined and the staff informed, the district administration should address the full range of training needs before the program is implemented. For example, inservice courses should be offered and sample materials distributed before teachers can be expected to begin embedding performance tasks in their classroom assessment plan.

8. Ensure comprehensive training

At the same time the specific training is taking place, steps to ensure continuing and comprehensive training in assessment should be taken. Teacher recruiters for the school district should increase their awareness of the preservice training in assessment that their applicants have received. When this training is not evident or adequate, the district should take steps to offer recertification and credit courses in testing and measurement and in classroom assessment. For those staff members unable to participate in the credit courses, alternative inservice plans should be written. Since teachers spend 25 to 35 percent of their time on assessment activities, the investment a district makes in this training should pay off in noticeable improvements in teacher performance.

9. Inform the public

One major step remaining prior to the implementation of the new program is to inform the parents and the community about the change. The message should emphasize the need for more and better data to permit improved decision making, with a constant focus on the goal of improving student achievement. More on public relations in Chapter 9.

10. Begin cautiously

This step could be called the shakedown cruise. The

program can begin, but only with a flexible implementation plan that assumes some things will go wrong. An occasional misunderstanding may develop and some changes may have to be made, but that should be considered a natural part of the process.

If all these steps are followed, the district should have an assessment program that:

- Delivers information that enables teachers to address the needs of individual students.
- Provides data that policy makers and curriculum developers can use to improve the overall quality of the districtwide instructional program.
- Is implemented by staff members who have been instrumental in its development, each of whom is fully knowledgeable about his or her role.
- Is accepted by the community as a true measure of their children's achievement and their school system's performance.

12-Step plan for test development

1. Selection of content

The first step in test development is to determine what content is to be covered by the test. Sometimes this is done by teachers and supervisors in a specific content area, like reading or math. In some cases, citizens may be polled to determine what content they want to see covered in a statewide testing initiative. In other cases, the school board and administration may have already established some priorities based on local or national concerns. Whoever makes the final selection of the content, this step must result in a list of the specific learning objectives to be assessed.

2. Development of test and item specifications

After the list of objectives has been completed, content and assessment specialists should prepare a document detailing how the content should be assessed. For example, if whole number division is to be assessed, the test developers must address such issues as these: "How many digits can the divisor and dividend have?" "Can there be remainders?" "What symbol will be used to represent division, the radical sign or the '÷'?"

The test and item specifications document should include:

- Sample items.
- Characteristics of the right answers.
- Characteristics of wrong answers.
- Specific words that may or may not be used in item construction.
- The number of items allocated to each objective.
- The readability level of the reading passages.

The major purpose of a test and item specifications document is to give detailed instructions to item writers so that items will be comparable in style and content even if they are generated by different people.

3. Item writing

After the specifications have been written and reviewed and determined to contain the necessary detail, it is time to develop the procedures for writing the test items. If a large number are to be written, it may be best to select the teachers and take them to a centralized retreat-type setting for as long as necessary to generate at least rough drafts of items. If fewer items are needed, after-school workshops can be used or the teachers can be trained so the items can be produced on a "take-home" basis. However the procedure is organized, the following steps must be completed:

- Item writers should be trained on the use of the test and item specifications guide. The reasons for each specification should be covered and all sample items explained.
- Item writers should choose the illustrations, diagrams, passages, and so forth, and begin constructing items.
- After a minimum number of items have been completed, the group should reconvene for a peer review. Questions that have arisen or problems that people encountered should be discussed by the group, and the leader should give directions for handling them.
- After this initial debriefing, periodic group discussions should be held to discuss new problems.
- The final product should be a set of draft items, at least 10 percent more than will be needed. Attachments should show the readability, the specific objective being measured, the grade level (if appropriate), and so forth.

4. Technical item review

The draft items must be reviewed by assessment specialists to determine how they conform to the test specifications. Readability levels should be recomputed; right answers must be clearly right and wrong answers must be clearly wrong.

5. Content item review

The items must then be reviewed by content specialists to determine if the content has been accurately portrayed.

6. Item bias review

When technically sound, content-valid items are ready, they should be turned over to a committee of specialists to determine if any potential bias exists toward males or females, ethnic or racial groups, handicapped individuals, or those from certain socioeconomic or geographic backgrounds. If an item can be rewritten to avoid the potential bias, it should be; if not, it should be discarded. (See below for more details on detecting bias.)

7. Legal item review

If these are considered high-stakes tests, a legal consultant may be asked to examine the items to make sure that nothing offensive or legally indefensible slips into the test.

8. Test form construction

After the pool of items is complete, test forms can be constructed. Attention must be given to items that may give away the answer to another item on the test, to the coverage of objectives (too few or too many items on one objective), and similar problems.

9. Pilot testing

Test forms should be administered to a group of potential examinees to determine the time needed, the administrative issues that must be covered, and so forth. This group need not be random, just typical.

10. Final construction of administrative directions

Although administrative steps should have been roughly drafted throughout the process, they must be finalized after pilot testing. Test administrators can be valuable editors of these documents.

11. Field testing

A random, representative sample of examinees should be given the test. It should be scored and the results examined carefully. Reliability and validity measures should be applied to determine the stability and accuracy of the measures. Any anomalies, problems, or inconsistencies should be addressed before beginning the routine examination schedule.

12. Implementation

After all the above steps have been followed, a sound test should be available to the schools.

Avoiding bias

No topic is as sensitive or can be as explosive as bias in tests. The possibility that one group may not have the same chance as others to succeed hits hard at our sense of ethics and fairness. Detecting bias and dealing with it, however, is not an easy task.

Many users of national standardized tests have praised the publishers for making strides recently in ridding their tests of obvious bias. But as noted earlier, there are still many concerns about hidden pitfalls in nationally standardized tests. The terrain is no less treacherous for local test developers.

Therefore it is advisable for a district to look carefully at *all* tests before they are included in its comprehensive assessment program. Here are some suggested steps for detecting bias in test items—both in purchased instruments and those that are being developed locally:

1. Conduct a field test

This is only one reason—but an important one—for a comprehensive field-testing of a new instrument. All elements of the school population must be included in the sample.

2. Sort the results

When the results of the pilot come back, a *statistical* measure should be applied to detect differentials in performance when students are sorted by race, ethnic group, gender, geographical location, socioeconomic level, handicapping condition, and so forth. This step merely flags differential performance. It does not explain *why* it happened.

3. Analyze items carefully

The reasons for differential performances may not be as obvious as it appears on the surface. As an exaggerated example: A test item is administered to a class consisting of 15 athletes and 15 computer hackers. If all of the athletes get the answer right and none of the hackers does, the assumption might be that the item is biased.

It is just as possible, however, that both groups should have known the answer, but their instructional programs were "biased," not the test. Just because one group misses certain items more than another does not warrant the immediate assumption that the measure is biased. It *may* be—but other aspects of the students' environment may have to be addressed instead of the test.

4. Apply judgment

Because of this limitation, a team of experts should review all test items. Since these individuals are asked to apply their judgment to the question, they must represent all of the ethnic and socioeconomic groups, genders, geographical regions, handicapping conditions and other categories that exist in the student body—including white males. These judges should examine each item for stereotyping, words that might have different connotations in different cultures, situations that may be foreign to some children because of their families' socioeconomic status, illustrations that may not be equally familiar to everyone being tested, and so forth.

5. Don't go overboard

The test developers and/or selectors should take the results of these reviews very seriously. But while it is important to be sensitive to such dangers as stereotyping—they should be careful not to go overboard. A question that creates a scenario in which all boys elect to take dance class while all girls take soccer can have as negative a repercussion as if it were the other way around.

6. Beware of geographical bias

With most of the emphasis being placed on avoiding bias because of race, gender or culture, another type sometimes slips through, especially where commercial tests are developed by people from other parts of the country. That is geographical bias: Taking for granted something integral to one region's culture but unknown, or different, in another.

A true example of geographical bias, as reported by testing specialists in Maryland, involved a national test—written by a West Coast publisher—that included a drawing of a store front with a sign that said, "Crabs—$3.99 each." The test writer, of course, was thinking of Alaskan king crabs. The purpose of the question was lost, however, as the students from the shores of the Chesapeake Bay tittered or guffawed over the price. Their parents were watermen, and the price of Maryland blue crabs was closer to $3.99 a bushel.

Staff development a necessity

Ideally, every teacher should be trained in assessment in college, long before he or she sets foot in a classroom. However, because there are tens of thousands of teachers already in the field without even a basic knowledge of the subject, this poses a major inservice problem for school districts. While a full-fledged staff development effort is the ultimate goal, there are ways a school district can begin to fill this gap without a major investment.

Evening courses

Many school systems have joined forces with a nearby college or university to offer evening courses for graduate credit that can be tailored specifically to the needs of classroom assessment. For example, the standard Tests and Measurement course, which often labors over statistics, can be modified to cover day-to-day assessment issues. And instead of reviewing all the standardized measures available from the test publishers, the course could concentrate on local needs, like the development of teacher-made tests.

Activities should include critiques of the participants' own tests, work with colleagues to develop banks of test items that can be used for designing customized tests, or practice using different assessment techniques to measure one learning objective.

Half-day programs

If a cooperative effort with a college is not possible, some well-structured half-day inservice programs can effectively communicate basic skills in testing.

The objectives of the inservice program must be clear. The presentation should cover the types of questions used in the classroom: true-false, multiple-choice, short answer, matching, or essay. Guidelines for the development of each type of question can be covered quickly and effortlessly if the instructor takes a practical, rather than scholarly, approach.

Small groups should construct examples of each type of question to provide some experience in writing test items. This, in turn, will generate more questions and concerns, discussion and analysis. A worksheet can be distributed illustrating potential flaws.

Throughout the entire training process, the emphasis must be on the need for test questions that measure something *important* in a *fair* and *consistent* manner. Students should be credited with a right answer if, and only if, they know the content that is being sampled by the question. Conversely, they should get it wrong if they do not know the content. By keeping this concept as the focal point of the entire course, the instructor will necessarily cover such critical issues as items with more than one right answer, bias toward different populations, and testing of trivial knowledge.

A workshop like this can cover only the very basics, but it could stimulate an interest in more elaborate concepts. In so doing, it might help demystify the test development process so that participants will become more aware—and demanding —about the quality of assessment instruments used in their own classrooms.

Code of Fair Testing Practices in Education

The Joint Committee on Testing Practices, a cooperative effort of several professional organizations, is devoted to the advancement of the quality of testing practices.

In 1988, the committee developed and published a document called the "Code of Fair Testing Practices in Education," which was meant to apply broadly to the use of tests in education (admission, educational assessment, educational diagnosis, and student placement). It was not designed to cover employment testing, licensure, or certification testing.

Although the Code is considered relevant to many educational tests, it is directed primarily at professionally developed tests, such as those sold by commercial test publishers or used in formally administered testing programs. It is not intended to cover tests made by individual teachers for use in their classrooms.

The Code addresses the role of test developers and test users separately. ***Test users*** are people who select tests, commission test development services, or make decisions on the basis of test scores. ***Test developers*** are people who actually construct tests as well as those who set policies for particular testing programs. The roles may overlap, as when a state education agency commissions test development services, sets policies that control the test development process, and makes decisions on the basis of test scores.

The Code presents standards for educational test developers and users in four areas:

A. Developing/Selecting Tests
B. Interpreting Scores
C. Striving for Fairness
D. Informing Test Takers

Organizations, institutions, and professionals who endorse the Code commit themselves to safeguarding the rights of the test takers by following the principles listed. It is intended to be consistent with the *Standards for Educational and Psychological Testing* published by the same group in 1985, except that it differs in both audience and purpose.

The Code is meant to be understood by the general public; it is limited to educational tests; and the primary focus is on those issues that affect the proper use of tests.

(The Joint Committee was initiated by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. In addition to those three groups, the American Association for Counseling and Development/Association for Measurement and Evaluation in Counseling and Development, and the American Speech-Language-Hearing Association are also sponsors of the Joint Committee.)

The Code, which follows verbatim, is not copyrighted. Reproduction and dissemination are encouraged by the Joint Committee, whose mailing address is: American Psychological Association, 1200 17th Street N.W., Washington, DC 20036.

A

Developing/ Selecting Appropriate Tests

Test developers should provide the information that test users need to select appropriate tests.

Test Developers Should:

1. Define what each test measures and what the test should be used for. Describe the population(s) for which the test is appropriate.
2. Accurately represent the characteristics, usefulness and limitations of tests for their intended purpose.
3. Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience(s).
4. Describe the process of test development. Explain how the content and skills to be tested were selected.
5. Provide evidence that the test meets its intended purpose(s).
6. Provide either representative samples or complete copies of test questions, directions, answer sheets, manuals and score reports to qualified users.
7. Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.
8. Identify and publish any specialized skills needed to administer each test and to interpret scores correctly.

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for the intended test-taking populations.

Test Users Should:

1. First define the purpose for testing and the population to be tested. Then select a test for that purpose and that population based on a thorough review of the available information.
2. Investigate potentially useful sources of information, in addition to test scores, to corroborate the information provided by tests.
3. Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.
4. Become familiar with how and when the test was developed and tried out.
5. Read independent evaluations of a test and of possible alternative measures. Look for evidence required to support the claims of test developers.
6. Examine specimen sets, disclosed tests

or samples of questions, directions, answer sheets, manuals and score reports before selecting a test.

7. Ascertain whether the test content and norms group(s) or comparison group(s) are appropriate for the intended test takers.
8. Select and use only those tests for which the skills needed to administer the tests and interpret scores correctly are available.

B

Interpreting Scores

Test developers should help users interpret scores correctly.

Test Developers Should:

9. Provide timely and easily understood score reports that describe test performance clearly and accurately. Also explain the meaning and limitations of reported scores.
10. Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.
11. Warn users to avoid specific, reasonably anticipated misuses of test scores.
12. Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to use such scores with the test.
13. Provide information that will help users gather evidence to show that the test is meeting its intended purpose(s).

Test users should interpret scores correctly.

Test Users Should:

9. Obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores.
10. Interpret scores taking into account any major difference between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.
11. Avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.
12. Explain how any passing scores were set and gather evidence to support the appropriateness of the scores.
13. Obtain evidence to help show that the test is meeting its intended purpose(s).

C

Striving for Fairness

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Developers Should:

14. Review and revise test questions and related materials to avoid potentially insensitive content and language.
15. Investigate the performance of test takers of different races, gender and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.
16. When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Users Should:

14. Evaluate the procedures used by test developers to avoid potentially insensitive content or language.
15. Review the performance of test takers of different races, gender and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences may have been caused by inappropriate characteristics of the test.
16. When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in the light of the modifications that were made.

D

Informing Test Takers

Under some circumstances, test developers have direct communication with test takers. Under other circumstances, test users communicate directly with test takers. Whichever group communicates directly with test takers should provide the information described below.

Test Developers or Test Users Should:

17. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether the test should be taken, or if an available alternative to the test should be used.
18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Strive to make such information equally available to all test takers.

Criteria for Evaluation of Proposed Student Assessment Systems

The National Forum on Assessment was established in 1990 by leaders of several education, civil rights, and advocacy groups in response to proposals by federal authorities for a national test. Its deliberations focused on the purposes of student assessment in general, and as a result the forum adopted the following paper in April, 1991. The paper, called "Criteria for Evaluation of Proposed Student Assessment Systems," was based on the belief that fundamental changes in assessment were called for, but tests may not be the answer —and especially not national tests.

"Because assessment affects educational standards, instructional methods, curricula, school structure, and governance, assessment decisions should not be made without consideration of these other factors," the introduction said. "To provide guidelines for evaluating proposed assessment programs at any level, we offer the following criteria:"

1.

Educational standards specifying what students should know and be able to do should be clearly defined before assessment procedures and exercises are developed.

For assessment information to be valid and useful, assessment must be based on a consensus definition of what students are expected to learn, and the expected level of performance, at various developmental stages. Such standards, which might also be called intellectual competencies, are not discrete pieces of information or isolated skills, but important abilities, such as the ability to solve various kinds of problems or to apply knowledge appropriately.

The standards should be determined through open discussion among subject-matter experts, educators, parents, and others, including those concerned with the relationship between school learning and life outside of school. Without a consensus on standards, there is little likelihood of a valid assessment.

2.

The primary purpose of the assessment system should be to assist both educators and policymakers to improve instruction and advance student learning.

Students, educators, parents, policymakers, and others have different needs for assessment and different uses for assessment information. For example, teachers, students, and their parents want information on individual achievement, while policymakers and the public want information for accountability purposes. In all cases, the system should be designed to provide not just numbers or ratings, but useful information on the particular abilities students have or have not developed.

All purposes and uses of assessment should be beneficial to students. If they cannot be shown to be beneficial, they should not be used at all.

3.

Assessment standards, tasks, procedures, and uses should be fair to all students.

Because individual assessment results often affect students' present situation and future opportunities, the assessment system, the standards on which it is based, and all its parts must treat students equitably. Assessment tasks and procedures must be sensitive to cultural, racial, class, and gender differences, and to disabilities, and must be valid for and not penalize any groups. To ensure fairness, students should have multiple opportunities to meet standards and should be able to meet them in different ways. No student's fate should depend upon a single test score.

4.

The assessment exercises or tasks should be valid and appropriate representations of the standards students are expected to achieve.

A sound assessment system provides information about the full range of knowledge and abilities considered valuable and important for students to learn, and therefore requires a variety of assessment methods. Multiple-choice tests, the type of assessment most commonly used at present, are inadequate to measure many of the most important educational outcomes, and do not allow for diversity in learning styles or cultural differences. More appropriate tools include portfolios, open-ended questions, extended reading and writing experiences which include rough drafts and revisions, individual and group projects, and exhibitions.

5.

Assessment results should be reported in the context of other relevant information.

Information about student performance should be one part of a system of multiple indicators of the quality of education. Multiple indicators permit educators and policymakers to examine the relationships among context factors (such as type of community, socioeconomic status of students, and school climate), resources (such as expenditures per student, physical plant, staffing, and money for materials and

equipment), programs and processes (such as curriculum, instructional methods, class size, rates, employment, and further education). Statements about educational quality should not be made without reference to this information. [More on this concept in Chapter 10 of this report.]

6.

Teachers should be involved in designing and using the assessment system.

For an assessment system to help improve learning outcomes, teachers must fully understand its purposes and procedures and must be committed to, and use, the standards on which it is based. Therefore, teachers should participate in the design, administration, scoring, and use of assessment tasks and exercises.

7.

Assessment procedures and results should be understandable to all concerned.

Assessment information should be in a form that is useful to those who need it — students, teachers, parents, legislators, employers, postsecondary institutions, and the general public. At present, test results are often reported in technical terms that are confusing and misleading, such as grade-level equivalents, stanines, and percentiles. Because everyone should be able to understand assessment results and the procedures on which they are based, the information must be readily comprehensible. [More on this in Chapter 9.]

8.

The assessment system should be subject to continuous review and improvement.

Large-scale, complex systems are rarely perfect, and even well-designed systems must be modified to adapt to changing conditions. Plans for the assessment system should provide for a continuing review process in which all concerned participate.

CHAPTER SEVEN

The Front Line of Assessment

A working knowledge of assessment techniques is critical if teachers are to use it effectively to help students, yet training in the field is woefully inadequate.

After all the talk about NRTs, CRTs, performances, and portfolios, the most essential assessment occurs every day in every classroom in our schools. The annual administration of standardized tests is meaningless in comparison. Even the end-of-semester exams are trivial next to it. The activities that determine grades on report cards, although closer, still don't compare.

The most essential measurement—the most important act of data-gathering—takes place dozens of times every day as a teacher asks questions, listens to answers, observes behaviors, and adapts his or her materials or technique to best teach each student.

It's the dynamic exchange in a language arts, social studies, or mathematics group that causes a teacher to think, "Why didn't Carrie volunteer that answer?" or "I didn't know Joseph was so interested in subtraction."

Natural but invisible

Observations and conclusions like these flow naturally, but often invisibly, out of classroom assessment activities. They can be so subtle that many teachers will underestimate their importance, assuming naively that more important information results from a "test." And beyond that, when they disagree with the conclusions reached by a norm-referenced test, either they doubt their own abilities or begin to develop a disdain, and often a fear, of the standardized measures.

In reality, it is through classroom assessment that attitudes, skills, knowledge, and thinking are fostered, nurtured, and accelerated—or stifled. What happens because of standardized tests pales in comparison with the earth-shaking effect these day-to-day assessments can have on the lives of children.

Thus it is essential that the definition of "assessment" be broadened beyond "tests" and even beyond the "alternative assessments" discussed in the previous chapter. When educators realize that "assessment" includes every activity that produces information about students, they open the door to a true revolution in classroom instruction.

Training is sparse

To be able to use all of those various forms of assessment effectively, as well as to be literate participants in the debate over current and future testing programs, teachers and administrators need a basic grasp of measurement and assessment concepts and methodology. Unfortunately, despite the current and future expectations for teachers and administrators, a majority of states require no training in testing, measurement, or assessment techniques for education majors. Meaningful inservice activities are isolated and superficial. Even those educators who choose on their own to take college courses in assessment are often faced with training in esoteric statistical concepts focusing almost entirely on traditional norm-referenced testing.

More and more pressure is being felt at the classroom and school level as the accountability juggernaut moves on. For their own survival, front-line educators must be able to explain their own testing practices as well as understand those being mandated by their states and their districts.

Even without considering the growing press for accountability and the advancements in testing, researchers estimate that teachers spend 30 to 50 percent of their class time engaged in assessment activities. These range from watching children interact to asking questions to administering end-of-year exams. In fact, any time a teacher gathers information and evaluates it to help determine how best to teach a student, it requires the manipulation of basic assessment concepts. But does the average teacher have the skills to do it effectively?

"Overcoming the neglect"

Richard J. Stiggins directs the Center for Classroom Assessment at the Northwest Regional Education Laboratory. He

specializes in training teachers in assessment, and his feeling about the situation is symbolized by the title of one of his papers: "Teacher Training in Assessment: Overcoming the Neglect."

Stiggins is outspoken in his criticism of teacher-training institutions for failing to address this need and of most states for failing to require such training for certification. Why the neglect? Stiggins offers five possible reasons:

1. There is a tendency in education to focus on process rather than outcomes. In an environment where process-oriented evaluations of students, teachers, and programs rule the day, training in the assessment of outcomes may not be regarded as central to the evaluation task.
2. These courses often have a reputation as being somewhat more academically demanding than typical education courses. Over the years, their unpopularity with students and faculty have led to their elimination from programs.
3. The systematic assessment of outcomes may be seen by school personnel as being too risky. If schools are very clear about their achievement targets and are clear and public about their assessments of those outcomes, there is always the chance that somebody in the community will disagree either with the target or the assessment.
4. Many educators assume that the quality of assessment in the classrooms is assured from outside the classroom. Even textbooks are now accompanied by their own quizzes, unit tests and computerized item test banks.
5. The assessment training offered in the past has traditionally failed to reflect any sensitivity whatever to the realities of the classroom.

Stiggins concludes that there is "compelling evidence that as a society we really don't care whether students learn anything in school or not." He will, of course, back down on that position, explaining that what he really means is that "we spend all available resources training teachers and administrators to produce learning, put that training to work in schools, and then allocate virtually no resources to train practitioners or interested others in sound methods of assessing the outcomes of our efforts."

Help is on the way

Many researchers have jumped into the breach by focusing on upgrading the skills of classroom assessment. For example, thanks to a concerted effort by the National Council on Measurement in Education (NCME) and many individuals and organizations associated with educational measurement, progress is being made toward updating the content of college education courses. Some of their materials are reproduced elsewhere in this report. These should be required reading for anybody involved either in evaluating student performance or making decisions based on the results. They should also form the basis for staff development programs within school districts to ensure that each teacher is well versed in the techniques and the pitfalls of classroom assessment.

A short course in classroom assessment

One of Stiggins' major concerns is that school districts have equated the concept of "assessment" with standardized tests and ignored the tremendous influence of a teacher's evaluation processes on a child's learning.

To emphasize the need for this kind of training, Stiggins stresses the many ways classroom assessment is used.

Teachers use it . . .

Teachers use assessment in the classroom for at least three purposes:

- To inform specific decisions, like diagnosing needs or grouping students for instruction.
- To teach.
- To control student behavior.

. . . students use it . . .

It is often overlooked, but students are even more important users of assessment results than their teachers:

- Students set personal expectations of academic excellence based on teachers' evaluation of prior achievement.
- Students decide whether they or someone else controls their academic well-being based on the relationship they see between how hard they work and how much the teacher's classroom assessments say they have learned.
- Students decide if, what, when, where, and with whom to study based on feedback from classroom assessments.

. . . and parents use it

Parents are also users of classroom assessments, making decisions critical to their children's academic well-being based on:

- The match between their expectations for their child and the child's apparent achievement.
- Their perceptions of the "teachability" of their child.

The purpose

Given all those critical uses, he asks, it is vital that the purpose of every assessment be clear. "What might happen

if a teacher was insensitive to how students are likely to use a particular piece of assessment information?" he asked. "What might happen if various assessment purposes became confused, such as when an assessment meant for diagnosis or practice became woven into the student's report card grade?"

The target

It must also be clear to teachers what achievement targets they want the students to hit:

- Knowledge—They expect students to master substantive, verbalizable subject matter knowledge, like science or history facts.
- Thinking—They often expect specific higher-order thinking operations or problem-solving skills to be demonstrated by students using that knowledge, such as in drawing inferences.
- Behaviors—Teachers expect achievement-related skills to be exhibited by the student in specific academic contexts, such as in giving a speech.
- Products—They expect achievement-related products to be created by the student, and they want those products to possess certain specific attributes, as in a research report.
- Affect—Teachers want certain motivation, attitudes, values, and appreciations to be expressed by the student in specific terms.

"For teachers to assess effectively (and to teach effectively)," he summarized, "they must possess a clear and highly refined vision of each of the targets they hold as valuable for their students. They must know in explicit terms what students are supposed to know, think, and be able to create when instruction is completed. And further, they must know how those targets translate into assessments that make sense."

The circumstances arise, he suggests, where teachers might aim for two or three or more of these targets at the same time. "If so, what does that suggest about the need for teachers, students, and parents to understand and be able to communicate about achievement targets?" he asked. And another question: "What proportion of the achievement targets we value as outcomes for students can be translated into objective multiple-choice test formats?"

Assessment is forever

For all of the available types of assessments to be entirely valid for each student, the teacher must be fully aware at all times of what he or she is doing, hearing, saying and remembering. Since there is a potential for change in any act of assessment, the teacher must first consider how to recognize when it is taking place.

While students are engaged in independent or group activities, whether reading or recess, teachers observe. At times, they document the behaviors they see with the intention of adding something to the instructional process.

Sometimes only the teachers' behavior must be structured. But often it is inconvenient to wait for certain skills or behaviors to occur in the normal course of events, so artificial circumstances are introduced. Once that takes place, the students are required to modify their behavior to demonstrate some skill, knowledge, or line of thinking.

Natural or artificial?

Assessments are neither one or the other; they are a combination of both. They fall on a continuum that ranges from naturally occurring to nonsensically artificial. An example of an assessment close to the natural end of the spectrum is one in which knowledge or understanding is used spontaneously during play. If a first grader writes a letter to Santa Claus without consulting parents or teachers, it reveals a great deal of information about writing skills, language development, and probably the impact of toy advertising.

The extremely artificial end of the continuum is, unhappily, often illustrated by standardized tests that require the reading of a passage especially constructed for the test with a specific "readability" and length. When a low-achieving fifth-grader is asked literal comprehension questions from a contrived passage about the Franco-Prussian War, the activity is approaching the "nonsensically artificial."

"The quality of instruction is a function of teachers' understanding of the strengths and weaknesses of their students. The depth of that understanding, in turn, hinges on the quality of teachers' assessment of student achievement."

—*Richard J. Stiggins*

Ideally, classroom assessment should be a rich sampling from this continuum that allows naturally occurring demonstrations of knowledge while also constructing situations that encourage the demonstration of understanding.

Variety needed

Particular emphasis should be given at this point to the need for variety in assessment. No teacher should give up the paper-and-pencil tests that have been used for years, nor should any teacher rely entirely on this type of measure.

Just as the accountability movement is calling for a variety of indicators of schooling, then in each classroom, on a smaller scale, teachers should look for a variety of indicators of learning and understanding for each student.

Kids are smart even if they aren't

Try a little humor in the workshop

Assessment workshops often have a deadly image. Even an advocate as strong as Richard Stiggins speculates that is one of the reasons people don't take them. So in offering local teacher workshops on assessment, it helps to use humor.

Everyone has an assessment horror story to share that can serve as the centerpiece for discussion and analysis. Teachers are often more responsive to workshops like these if they are asked to approach them as lifetime test-takers, rather than test-makers.

You know siskum . . .

"Pop quizzes" can break the ice, especially when they prove a point. Like with this question:

> The results of the test showed that the end products of siskum are:
> A. bilge and sup
> B. seep
> C. smog
> D. bilge

It doesn't take long for most of the participants to realize that it's a joke. But is it?

The instructor would point out that a "smart" student would pick the right answer: A, because it's the only plural answer and the question asked for "products." Nonsense tests like these often get a chuckle—and an understanding of how students can get items right when they know nothing of the content.

True or false?

Bright students are also wise to questions that use the words "always" and "never," and have a strategy on true-false tests that works more often than not: Everyone knows, the kids will reveal, that the first one is always true. The second one is false because the examiner doesn't want us to think they're all true. The third is false so we don't think they'll go true, false, true, false. And so on.

The moral of this activity is not to make it easy for the students to get an item right if they don't know the answer.

Everyday classroom assessment tools

When it comes to making all of this happen, there are no deep, dark secrets of the ages known by only an elite few. In fact, all of the types of alternative assessments discussed so far in this report can be constructed by classroom teachers and used both to improve instruction and to inform parents and administrators of student progress.

Following is a brief discussion of some of these assessment tools:

Asking questions is perhaps the oldest form of assessment and may still be the best. The teacher can set high expectations with some well-chosen questions to spur students into creative and critical thinking, or the teacher can give the impression that only low-level rote learning is important. Teachers can enhance thinking skills simply by giving students a chance to think. Too many teachers are impatient and allow only a couple of seconds for students to respond before calling on someone else.

The importance of asking the right question was illustrated once in a graduate course in reading remediation. The instructor read aloud the Truman Capote story, "A Christmas Memory," during the last class before the holiday break. The weather was snowy, anticipation ran high over vacation and travel plans, and the instructor was challenged to practice what he preached. The beautiful story was well read and added warmth to already glowing holiday spirit.

At the end of the story everything was silent, each student pondering the personal effect of the experience. Then the instructor said, "Let me read a comprehension question for you. Why was the baby buggy used?"

The impact on the class was instantaneous. Most of them didn't remember a baby buggy. Their overwhelming reaction was, "What does that have to do with this story?"

It was a literal comprehension question, like many that are constructed and used every day in every classroom, but it did *not* have anything to do with the story.

A question like, "Have you ever had a relationship like the one between the boy and the old woman?" could have inspired the class to find a place for this piece of literature in their own lives. It was a powerful lesson on remedial reading instruction —and on the impact of a question.

Oral questioning can also be used for recall of facts, as frequently demonstrated in classrooms. Spelling bees or similar situations that require a rapid response to an oral mathematics problem are excellent examples.

Once again, not all questioning should be of the same type. It all depends on the kind of knowledge that is being assessed. If a teacher wants to find out if the students know their multiplication facts, one kind of question will be best. If instead, he or she wants to know if the students can

apply multiplication facts in a real-life setting, then another type of question may be more appropriate.

Observations of student behaviors color teachers' attitudes and opinions about the understanding level of each student. If there is a system for documenting these observations, then enormous power can be added to the evaluation of progress.

Many successful teachers create a grid displaying each child's name down the left side and the behaviors they wish to observe across the top (or vice versa). As they move about the classroom during independent or group work, they check the appropriate box next to the child's name. These can be notations about whether the student knows the material or not, or whether the child is applying critical thinking to the question or just giving a rote answer. The chart can also note the children's nonverbal behavior: Are they bored, fidgety, or comfortable, and do they stay on task? It may take several days or a week to observe all of the necessary behaviors for each child, but the information is well worth the wait.

Self-evaluation. A teacher can gain a lot of insight into the progress of students simply by asking them how they are doing — whether they understand the work and where they need further help. This can be done verbally or in writing, and it can also be "standardized" through a checklist on which students periodically check off, fill in, or circle those things they do in relation to a certain subject.

Paper-and-pencil testing is and should be an integral part of classroom assessment. But the teacher should devote a lot of care to the construction of the tests to differentiate between those that will call for rote recall of facts and those designed to demonstrate higher-order thinking. Too often teacher-made tests emphasize the rote end of this range.

Most paper-and-pencil tests in the classroom are criterion-referenced. A unit of information has been taught and the most important aspects of that unit are tested. The purpose is not to compare the students' scores, but rather to determine the mastery or success level of each student.

Teachers can choose from among true-false, multiple-choice, completion, short answer and essay questions as they construct their tests, but they should vary formats depending on what they are trying to assess.

Some content areas and topics lend themselves more easily to one format than another, but creative teachers use those that are less typical. Classically, mathematics has been tested in a multiple-choice or completion format. However, asking a high school statistics class, "Why must we test the null hypothesis?" or a fifth grader "What does multiplication mean?" and requiring a written response could reveal more about the students' understanding than a page full of calculations.

In constructing tests, teachers should be continually careful about items that can put some students at a disadvantage (as described in the section on bias above), and to watch out for potential flaws in each item. A favorite in many teacher-training workshops is this true-false math item, which seems obvious on its face:

T F The sum of the angles of a triangle is 180 degrees.

Is this true or false? It depends. Is this Euclidian or non-Euclidian geometry? Does it mean interior or exterior angles? In other words, even seemingly simple items can be tricky.

Performance tasks

These alternative assessment tools available to classroom teachers were covered in detail in previous chapters, along with sample activities that could be adapted to individual classrooms. They give teachers greater opportunities to observe more complex thinking and understanding than many of the traditional methods used in America's classrooms.

Here is an example of that approach in an unlikely place —a first-grade classroom. The student, when asked to enter the school's science fair, decided to study whether plants would grow better if she sang to them. She planted seeds (control and experimental groups, of course) and gave them the same amount of light and water. One group she sang to and the other she did not.

At the end of several weeks of singing, she made a final measurement and found that the group that she had sung to was smaller than the control group. Her conclusion at the science fair was that singing does not help growth—unless, of course, she had chosen the wrong songs.

The point is that this one activity required curiosity, invention, horticultural concepts, patience, writing skills, and scientific thought. While she didn't win a science fair prize, she certainly won the prize for the integration of skills and the encouragement of experimentation.

Such assignments should be given regularly to children of all ages. Too often projects requiring sustained attention, multiple steps and integrated skills are reserved for older students.

Measuring thinking skills

The emphasis in this chapter has been on variety. Classroom assessment calls for many different tools to gather information about students. Although the testing of rote recall is at times necessary, it should never be the predominant purpose for assessment or the students will begin to direct their learning toward facts alone and away from thinking.

Stiggins, Evelyn Rubel and Edys Quellmalz in *Measuring*

Thinking Skills in the Classroom (a National Education Association publication) develop a classification system and encourage teachers to use it to ensure that the full range of thinking skills is covered in each assessment activity.

They categorize the five basic thinking skills as:

- recall
- analysis
- comparison
- inference
- evaluation.

They propose that teachers construct an Assessment Planning Chart for each unit they teach, with those five skills down the side and three columns across the top representing three common forms of classroom assessment: "Oral," "Test" and "Performance." They are to write in each box in the resulting grid at least one question or activity designed to demonstrate the child's mastery of each of the five thinking skills.

Examples of that grid appear later in this chapter. The book also contains a suggestion on using certain words as "trigger or action verbs" to use when asking the questions:

If you want to measure:	Use these key words in the exercise:		Illustration
Recall	define identify label list name	repeat what when who	List the names of the main characters in the story.
Analysis	subdivide break down separate	categorize sort	Break the story down into different parts.
Comparison	compare contrast	differentiate distinguish	Compare the themes of these two stories.
Inference	deduce predict infer speculate	anticipate what if . . . apply conclude	How might we make this character more believable?
Evaluation	evaluate judge assess appraise defend	argue recommend debate why critique	Evaluate this story. Is it well written? Why or why not?

The authors feel this effort is especially critical at this juncture because educators are giving higher priority to the development of students' thinking skills. This has pointed up not only the inadequacy of instruction but also the inadequacy of the assessment of those skills. They insist that the situation must change: "As educators, we want students to do far more than just restate facts we have taught them. Our goal is to teach students to *use* the information at their disposal. We want them to think. Very simply, we are doomed to fall far short of this goal if we are unable to monitor development and measure student mastery of the thinking skills we need to impart. If we cannot

measure the extent to which a skill has been mastered, we cannot determine what to teach next. If we cannot measure the skills we teach, we cannot know if instruction is effective."

Teacher education the key

If substantive progress is to be made in relating assessment to instructional improvement, the process cannot be allowed to be hit-or-miss.

Robert J. Swartz and D.N. Perkins wrote in 1989, "We encourage school administrators seriously interested in the successful teaching of thinking to support their teachers' efforts by providing the personnel and expertise needed for formal summative evaluations. . . . The beneficial results will be better local decision-making and a growing pool of experience from school to school about effective ways of teaching thinking."

And if we can reach the stage where these competencies are commonplace in teacher education programs, where states expect them of all certificated employees, and where they are regularly used in our classrooms, then the quality of assessment will surely increase.

The most important thing a district superintendent can do in the area of assessment, Stiggins asserted, is to make sure his or her own skills in assessment are adequate, and then to make sure that the skills of the administrative and instructional staffs are adequate as well.

"The power will rest with the knowledgeable," he said.

Standards for Teacher Competence in Educational Assessment of Students

These standards were developed by a committee of experts appointed by three collaborating educational associations: the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association "out of concern that the potential educational benefits of student assessments be fully realized." In early 1991, the same groups embarked on a similar venture, in cooperation with the American Association of School Administrators, the National Association of Elementary School Principals, and the National Association of Secondary School Principals to develop a similar paper describing the competencies needed by school administrators for interacting with teachers, patrons, and the central office in terms of understanding and using student assessment.

The following section is excerpted from a paper published by the sponsoring organizations in 1990 of which "reproduction and dissemination are encouraged."

A standard is defined here as a principle generally accepted by the professional associations responsible for this document. Assessment is defined as the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weaknesses, to judge instructional effectiveness and curricular adequacy, and to inform policy.

The various assessment techniques include, but are not limited to, formal and informal observation, qualitative analysis of pupil performance and products, paper-and-pencil tests, oral questioning, and analysis of student records. The assessment competencies included here are the knowledge and skills critical to a teacher's role as an educator. It is understood that there are many competencies beyond assessment competencies which teachers must possess.

By establishing standards for teacher competence in student assessment, the associations subscribe to the view that student assessment is an essential part of teaching and that good teaching cannot exist without good student assessment. Training to develop the competencies covered in the standards should be an integral part of preservice preparation. Further, such assessment training should be widely available to practicing teachers through staff development programs at the district and building levels.

The standards are intended for use as:

- a guide for teacher educators as they

Continued on page 70

Continued from page 69

design and approve programs for teacher preparation
- a self-assessment guide for teachers in identifying their needs for professional development in student assessment
- a guide for workshop instructors as they design professional development experiences for in-service teachers.
- an impetus for educational measurement specialists and teacher trainers to conceptualize student assessment more broadly than has been the case in the past.

The scope of a teacher's professional role and responsibilities for student assessment

There are seven standards in this document. In recognizing the critical need to revitalize classroom assessment, some standards focus on classroom-based competencies. Because of teachers' growing roles in education and policy decisions beyond the classroom, other standards address assessment competencies underlying teacher participation in decisions related to assessment at the school, district, state and national levels.

The scope of a teacher's professional role and responsibilities for student assessment may be described in terms of the following activities. These activities imply that teachers need competence in student assessment and sufficient time and resources to complete them in a professional manner.

Activities Occurring Prior to Instruction

(a) Understanding students' cultural backgrounds, interests, skills, and abilities as they apply across a range of learning domains and/or subject areas; (b) understanding student motivations and their interests in specific class content; (c) clarifying and stimulating the performance outcomes expected of pupils; and (d) planning instruction for individuals or groups of students.

Activities Occurring During Instruction

(a) Monitoring pupil progress toward instructional goals; (b) identifying gains and difficulties pupils are experiencing in learning and performing; (c) adjusting instruction; (d) giving contingent, specific and credible praise and feedback; (e) motivating students to learn; and (f) judging the extent of pupil attainment of instructional outcomes.

Activities Occurring After the Appropriate Instructional Segment (lesson, class, semester, grade)

(a) Describing the extent to which each pupil has attained both short- and long-term instructional goals; (b) communicating strengths and weaknesses based on assessment results to students, and parents or guardians; (c) recording and reporting assessment results for school-level analysis, evaluation and decision-making; (d) analyzing assessment information gathered before and during instruction to understand each student's progress to date and to inform future instructional planning; (e) evaluating the effectiveness of instruction; and (f) evaluating the effectiveness of the curriculum and materials in use.

Activities Associated with a Teacher's Involvement in School Building and School District Decision-Making

(a) Serving on a school or district committee examining the school's and district's strengths and weaknesses in the development of its students; (b) working on the development or selection of assessment methods for school building or school district use; (c) evaluating school district curriculum; and (d) other related activities.

Activities Associated with a Teacher's Involvement in a Wider Community of Educators

(a) Serving on state committees asked to develop learning goals and associated assessment methods; (b) participating in reviews of the appropriateness of district, state or national student goals and associated investment methods; and (c) interpreting the results of state and national student assessment programs.

Each standard that follows is an expectation for assessment knowledge or skill that a teacher should possess in order to perform well in the five areas just described. As a set, the standards call on teachers to demonstrate skill at selecting, developing, applying, using, communicating, and evaluating student assessment information and student assessment practices. A brief rationale and illustrative behaviors follow each standard.

STANDARDS

1

Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

Skills in choosing appropriate, useful, administratively convenient, technically adequate, and fair assessment methods are prerequisite to good use of information to support instructional decisions. Teachers need to be well-acquainted with the kinds of information provide by a broad range of assessment alternatives and their strengths and weaknesses. In particular, they should be familiar with the criteria for evaluating assessment methods in light of instructional plans.

Teachers who meet this standard will have the conceptual and application skills that follow.

They will be able to use the concepts of assessment error and validity when developing or selecting their approaches to classroom assessment of students. They will understand how valid assessment data can support instructional activities such as providing appropriate feedback to students, diagnosing group and individual learning needs, planning for individualized educational programs, motivating students, and evaluating instructional procedures.

They will understand how invalid information can affect instructional decisions about students. They will also be able to use and evaluate assessment options available to them, considering among other things, the cultural, social, economic and language backgrounds of students. They

Continued on next page

will be aware that different assessment approaches can be incompatible with certain instructional goals and may impact quite differently on their teaching.

Teachers will know, for each assessment approach they use, its appropriateness for making decisions about their pupils. Moreover, teachers will know of where to find information about and/or reviews of various assessment methods. Assessment options are diverse and include text- and curriculum-embedded questions and tests, standardized criterion-referenced and norm-referenced tests, oral questioning, spontaneous and structured performance assessments, portfolios, exhibitions, demonstrations, rating scales, writing samples, paper-and-pencil tests, seatwork and homework, peer- and self-assessments, student records, observations, questionnaires, interviews, projects, products, and others' opinions.

2

Teachers should be skilled in *developing* assessment methods appropriate for instructional decisions.

While teachers often used published or other external assessment tools, the bulk of the assessment information they use for decision-making comes from approaches they create and implement. Indeed, the assessment demands of the classroom go well beyond readily available instruments.

Teachers who meet this standard will have the conceptual and application skills that follow.

Teachers will be skilled in planning the collection of information that facilitates the decisions they will make. They will know and follow appropriate principles for developing and using assessment methods in their teaching, avoiding common pitfalls in student assessment. Such techniques may include several of the options listed at the end of the first standard. The teacher will select the techniques which are appropriate to the intent of the teacher's instruction.

Teachers meeting this standard will also be skilled in using student data to analyze the quality of each assessment technique they use. Since most teachers do not have access to assessment specialists, they must be prepared to do these analyses themselves.

3

The teacher should be skilled in administering, scoring and interpreting the results of both externally produced and teacher-produced assessment methods.

It is not enough that teachers are able to select and develop good assessment methods; they must also be able to apply them properly. Teachers should be skilled in administering, scoring and interpreting results from diverse assessment methods.

Teachers who meet this standard will have the conceptual and applications skills that follow.

They will be skilled in interpreting informal and formal teacher-produced assessment results, including pupils' performances in class and on homework assignments. Teachers will be able to use guides for scoring essay questions and projects, stencils for scoring response-choice questions, and scales for rating performance assessments. They will be able to use these in ways that produce consistent results.

Teachers will be able to administer standardized achievement tests and be able to interpret the commonly reported scores: percentile ranks, percentile band scores, standard scores and grade equivalents. They will have a conceptual understanding of the summary indexes commonly reported with assessment results: measures of central tendency, dispersion, relationships, reliability, and errors of measurement.

Teachers will be able to apply these concepts of score and summary indices in ways that enhance their use of the assessments that they develop. They will be able to analyze assessment results to identify pupils' strengths and errors. If they get inconsistent results, they will seek other explanations for the discrepancy or other data to attempt to resolve the uncertainty before arriving at a decision. They will be able to use assessment methods in ways that encourage students' educational development and that do not inappropriately increase students' anxiety levels.

4

Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

Assessment results are used to make educational decisions at several levels: in the classroom about students, in the community about a school and a school district, and in society, generally, about the purposes and outcomes of the educational enterprise. Teachers play a vital role when participating in decision-making at each of these levels and must be able to use assessment results effectively.

Teachers who meet this standard will have the conceptual and application skills that follow.

They will be able to use accumulated assessment information to organize a sound instructional plan for facilitating students' educational development. When using assessment results to plan and/or evaluate instruction and curriculum, teachers will interpret the results correctly and avoid common misinterpretations, such as basing decisions on scores that lack curriculum validity. They will be informed about the results of local, regional, state, and national assessments and about their appropriate use for pupil, classroom, school, district, state, and national education improvement.

5

Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.

Grading students is an important part of professional practice for teachers. Grading is defined as indicating both a student's level of performance and a teacher's valuing of that performance. The principles for using assessments to obtain valid grades are known and teachers should employ them.

Teachers who meet this standard will have the conceptual and application skills that follow.

They will be able to devise, implement, and explain a procedure for developing grades composed of marks from various

Continued on page 72

assignments, projects, in-class activities, quizzes, tests, and/or other assessments that they may use. Teachers will understand and be able to articulate why the grades they assign are rational, justified, and fair, acknowledging that such grades reflect their preferences and judgments.

Teachers will be able to recognize and to avoid faulty grading procedures such as using grades as punishment. They will be able to modify their grading procedures in order to improve the validity of the interpretations made from them about students' attainments.

6

Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

Teachers must routinely report assessment results to students and to parents or guardians. In addition, they are frequently asked to report or to discuss assessment results with other educators and with diverse lay audiences. If the results are not communicated effectively, they may be misused or not used. To communicate effectively with others on matters of student assessment, teachers must be able to use assessment terminology appropriately and must be able to articulate the meaning, limitations, and implications of assessment results.

Furthermore, teachers will sometimes be in a position that will require them to defend their own assessment procedures and their interpretations of them. At other times, teachers may need to help the public to interpret assessment results appropriately.

Teachers who meet this standard will have the conceptual and application skills that follow.

Teachers will understand and be able to give appropriate explanations of how the interpretation of student assessments must be moderated by the student's socioeconomic, cultural, language, and other background factors. Teachers will be able to explain that assessment results do not imply that such background factors limit a student's ultimate educational development. They will be able to communicate to students and to their parents or guardians how they may assess the student's educational progress.

Teachers will understand and be able to explain the importance of taking measurement errors into account when using assessments to make decisions about individual students. Teachers will be able to explain the limitations of different informal and formal assessment methods. They will be able to explain printed reports of the results of pupil assessments at the classroom, school district, state and national levels.

7

Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

Fairness, the rights of all concerned, and professional ethical behavior must undergird all student assessment activities, from the initial planning for and gathering of information to the interpretation, use, and communication of the results. Teachers must be well-versed in their own ethical and legal responsibilities in assessment.

In addition, they should also attempt to have the inappropriate assessment practices of others discontinued whenever they are encountered. Teachers should also participate with the wider educational community in defining the limits of appropriate professional behavior in assessment.

Teachers who meet this standard will have the conceptual and application skills that follow.

They will know those laws and case decisions which affect their classroom, school district, and state assessment practices. Teachers will be aware that various assessment procedures can be misused or overused resulting in harmful consequences such as embarrassing students, violating a student's right to confidentiality, and inappropriately using students' standardized achievement test scores to measure teaching effectiveness.

"There are one-story intellects, two-story intellects, and three-story intellects with skylights. All fact collectors who have no aim beyond their facts are one-story people. Two-story people compare, reason, generalize, using the labor of the fact collectors as their own. Three-story people idealize, imagine, predict—their best illumination comes from above through the skylight."

—*Oliver Wendell Holmes*

These are samples of assessment planning charts devised by Stiggins, Rubel and Quellmaz to measure thinking skills.

ASSESSMENT PLANNING CHART

GRADE LEVEL Elementary SUBJECT English TOPIC Poetry

	ORAL	TEST	PERFORMANCE
RECALL	Let's see if we can tell what poetry is.	A haiku has _____ lines and _____ syllables.	Compare some work of local poets. Choose the one you like best. As you know, artists are often not "recognized" until after their deaths so they often live in poverty. You serve on a board which could help this poet you have chosen. How would you go about convincing the board to give funds to this particular poet? Prepare your materials for convincing. Try them out on the class.
ANALYSIS	Let's see how many different kinds of poetry we can list.	Tell the purpose of an acrostic in poetry and write a brief example.	
COMPARISON	Let's compare several kinds of poetry as to rhyming technique, rhythm, structure, etc. Shall we first compare a limerick and a couplet?	Put an H if a haiku, a C if a couplet, and a D if a diamente. ___ a. Has 17 syllables ___ c. Has 7 lines ___ b. Has 2 lines ___ d. Has 3 lines.	
INFERENCE	Have several poems ready for use on the overhead, written on the board, or prepared as handouts. Students choose a poem and imagine how it might sound written as another form. Put examples on the board.	Rewrite this haiku as a limerick.	
EVALUATION	Do you think it's important to encourage the writing of poetry? Why or why not?	Some poetry is very structured. Do you prefer it to the less structured? Why or why not? Cite two examples of the kind of poetry you prefer.	

ASSESSMENT PLANNING CHART

GRADE LEVEL Junior High SUBJECT Social Studies TOPIC Electoral College

	ORAL	TEST	PERFORMANCE
RECALL	What is the electoral college?	As a member of the electoral college, you must vote: a. According to your own judgment b. As your constituency voted c. As the party tells you d. Only if you wish to do so.	Assume you're a U.S. senator. Propose a constitutional amendment that would make the popular vote the sole criterion for electing a president. Your amendment would do away with the electoral college. Prepare a speech to Congress defending your amendment. Be sure to— a. Analyze all elements of the issue, b. Compare elections with and without the college, c. Show how the voters are likely to react, and d. State and defend your values. (Or conduct a simulated debate on the Senate floor.)
ANALYSIS	How does the electoral college work?	Analyze the steps in the presidential election process, showing where the electoral college comes into play.	
COMPARISON	How do the social conditions that existed when the electoral college was formed differ from conditions now?	What is meant by the election theme "one person, one vote," and how does that relate to the electoral college?	
INFERENCE	If you were a presidential candidate elected by popular vote, could you still lose the election? How?	In which state is the electorate likely to oppose the use of the electoral college: a. California b. Illinois c. Etc.	
EVALUATION	Should the electoral college be abolished? Why or why not?	Which of the following is the best reason for maintaining the electoral college? a. Tradition. b. Fairness to large states. c. Etc.	

From Measuring Thinking Skills in the Classroom, revised edition, copyright 1988, National Education Association. Reprinted with permission.

76 BEST COPY AVAILABLE

ASSESSMENT PLANNING CHART

GRADE LEVEL High School SUBJECT Social Studies TOPIC Branches of Government

	ORAL	TEST	PERFORMANCE
RECALL	Who's in charge of the executive branch of government? The judicial branch? The legislature?		Set the class up as a democracy with three branches of government. Go through the actual process of passing a law. Determine if students know and can carry out each function. Strive to demonstrate the interactions among branches.
ANALYSIS		What can the executive branch do about an unfair law?	
COMPARISON	In America we have Congress. In Canada their legislature is called Parliament. How do they differ in structure? *OR* Compare the structure of the federal government with that of your state government.		
INFERENCE		You've decided motorcycles cause too many accidents. You'd like them banned from federally funded highways. To which government branch would you first appeal? a. Executive b. Judicial c. Legislative	
EVALUATION	Which branch of the government is the most important? Why?		

From Measuring Thinking Skills in the Classroom, revised edition, copyright 1988, National Education Association. Reprinted with permission.

This is one of the checklists provided to teachers in workshops in Fairfax County, Virginia, to help them with their classroom assessment activities. This is one to hand out to students to help them assess their own writing.

Self-assessment of Writing
Student Checklist

Things I do to plan my writing:			
• Drawing	Usually	Sometimes	Not usually
• Webbing	Usually	Sometimes	Not usually
• Data-retrieval charts	Usually	Sometimes	Not usually
• Discussion with others	Usually	Sometimes	Not usually
• Reading	Usually	Sometimes	Not usually
• Other:	Usually	Sometimes	Not usually
•	Usually	Sometimes	Not usually

Things I do to revise my writing:			
• Reread to see if it makes sense	Usually	Sometimes	Not usually
• Check for a beginning and an end	Usually	Sometimes	Not usually
• Add information	Usually	Sometimes	Not usually
• Take out information	Usually	Sometimes	Not usually
• Reorganize ideas	Usually	Sometimes	Not usually
• Carefully choose the best word	Usually	Sometimes	Not usually

Things I look at when I edit my writing:			
• Length of the sentences	Usually	Sometimes	Not usually
• Paragraphs	Usually	Sometimes	Not usually
• Quotation marks	Usually	Sometimes	Not usually
• Capital letters at the beginning of sentences	Usually	Sometimes	Not usually
• Capital letters on names	Usually	Sometimes	Not usually
• Commas	Usually	Sometimes	Not usually
• Punctuation at end of sentence	Usually	Sometimes	Not usually
• Spelling	Usually	Sometimes	Not usually
• Other:	Usually	Sometimes	Not usually

Other thoughts about my writing:

Guidelines for Test Item Writing

Here are some tips for teachers in putting together various types of tests. They have been culled from the materials used by assessment specialists in teacher-training workshops.

True-false

- The items should measure important objectives.
- Items should be stated positively; never use double negatives.
- Make each statement unequivocally true or false.
- Use an even mix of true and false statements—about half and half.
- Keep true and false items about the same length.
- Include only one concept in a statement.

Matching

- Use homogeneous options and items.
- Have more options than items.
- Use brief lists; place shorter words or phrases at the right.
- List options logically, alphabetically, or numerically.
- Describe the basis for matching and the number of times an option may be used to make a match.
- Scatter the correct and incorrect options in the list.
- Place all items and options on one page.

Multiple choice

- Present a self-contained question or problem in the stem.
- Include as much item content as possible in the stem.
- Avoid negatively stated items.
- Have only one correct or best answer.
- Make each option grammatically consistent with the stem.
- Make all options plausible and of similar length.
- Avoid options like "none of the above" or "all of the above."
- Avoid overlapping options.
- Randomly use each alternative position as the correct answer.
- Use four to five options to minimize guessing.

Short answer and completion

- Use a direct question rather than an incomplete statement.
- Write items that clearly imply the type of response desired.
- Place the blank at the end of the item or in the margin.
- Use only one, or at most two, blanks in one item.
- Avoid specific determiners (e.g., the word "an" to indicate the right answer begins with a vowel).
- Provide sufficient answer space.
- Make all blanks equal in length.
- Avoid using statements directly from the text.

Measure complex objectives

- The objective should permit the measurement of understanding.
- Construct items in a form different from that in which the material was originally presented.
- Use novel graphic or pictorial materials to measure principles that require students to apply knowledge.
- Use analogies to measure relationships, terms, or functions.
- Have students identify assumptions and analyze criteria.
- Have students discover relationships among similar topics.
- Have students select samples of principles or concepts.
- Provide for a condition contrary to fact (e.g., What might have happened if had not occurred?).
- Use various forms of charts, tables, statistics, etc.

Essay

- Structure the task.
- Specify limitations and explain each item's value.
- Avoid using optional questions.
- Prepare a trial response to the question.
- Construct a tentative scoring key in advance.
- Score all responses to a single question before scoring the next question.
- Decide in advance on how spelling, punctuation, and penmanship will be weighted.
- Score responses anonymously.

Guidelines for Administering a Performance Task

The administration of performance tasks has not been refined to the extent that guidelines are widely available to help teachers create situations and direct the students. The California Assessment Program has prepared a paper to help teachers develop writing prompts for the eight different types of writing it assesses. With thanks to Barbara Weiss of the CAP, the following guidelines are culled from that paper's section on *Report of Information*:

- Identify a subject for the report and describe it for the students in one to four sentences.
- The subject should be general, leaving to the student the choice of a specific subject and what information to include. For example, students might be asked to describe one of their classes for high school students in Australia, but without suggesting to students which class to choose or what to include in the report.
- Ensure that students readily recognize the subject and have an adequate context for reporting on it.
- Be sure to use the word "report." If students are to report on a process (how to do something, like how to find a job) or activity (skateboarding), help them understand that their purpose is not to give directions, but rather to increase readers' understanding of the subject and their interest in it.
- Identify who is asking the student to write the report (perhaps the student's teacher), or who will publish it (perhaps the school newspaper) or who will read it (perhaps other students).
- Use such words as "inform," "informative," or "information," and possibly "describe," "present," or "explain." The words "specific," "detailed," or "interesting" may be appropriate.
- Conclude the directions with a sentence that reiterates what readers will be expecting from the report.

CHAPTER EIGHT

The Politics of Assessment

Charges that the achievement of the nation's students has not improved in 20 years have provoked a controversy at the national level involving the threat of national standards and a national test.

Even as states and local school districts grapple with questions about the validity and relevance of their local testing programs, in Washington, D.C., the race to establish—or expand—national testing of America's school children threatened to become the hottest game in town.

Proponents argued that in light of evidence that student achievement has shown little improvement over the past 20 years, a national test is needed so that policymakers and parents can keep an eye on the big picture of student performance.

Opponents counter that students already are "tested to death" and that national exams—and especially national standards—will exert so much pressure on local curricula and teaching practices that they will do more harm than good. Furthermore, they argue, education is the prerogative of the states and should remain so.

A dizzying array of groups has pulled up chairs to the playing table. Here's a scorecard of the major players—and their positions—in "The National Assessment/Goals Game":

- ***National Education Goals Panel***—This group is the "designated hitter" established to follow through on the six national education goals established by the governors and President Bush in September 1989 at the Education Summit in Charlottesville, Virginia. The group was scheduled to issue its first "report card" on progress toward meeting the goals in September 1991.
- ***National Assessment of Educational Progress (NAEP)***—A Congressionally mandated project created 20 years ago to measure the scholastic achievement of students in elementary, middle, and high school. During its lifetime, it has issued more than 200 reports spanning 11 instructional areas. The contract is awarded by the National Center for Education Statistics in the U.S. Department of Education. It is currently administered by the Educational Testing Service, which also runs the Scholastic Aptitude Test (SAT).
- ***National Assessment Governing Board (NAGB)***—Created by Congress in 1988, the NAGB sets overall policy for the National Assessment of Educational Progress. While NAEP has been testing students for 20 years, its new governing board is increasing the frequency of some assessments and setting standards on what constitutes "basic," "proficient," and "advanced" performance on them.
- ***Council of Chief State School Officers (CCSSO)***—This organization, often referred to as the "Chiefs," represents the top education official in each state, usually an elected superintendent or an appointed commissioner. Its State Assessment Center releases an annual report of state education indicators containing state comparisons of factors such as Carnegie units required for graduation, the kinds of effective schooling initiatives underway in each state, plus demographic and fiscal background information.
- ***National Governors Association (NGA)***—This organization worked closely with the White House in launching the national goals effort. A task force on education developed the national goals and created the National Education Goals Panel, which became a separate entity. The NGA has contributed to the national debate by issuing four annual "Results in Education" reports in 1987-90 that contained blank columns for such outcome measures as student achievement and high school drop-out rate. The group's education department works with states to help reach the goals by providing technical assistance and developing policy guides and issue-oriented reports.
- ***Education Commission of the States (ECS)***—ECS is an interstate compact of 49 states and U.S. territories that studies policy issues to assist state leaders in improving the quality of education. The chairman is a governor and its primary constituencies are governors, legislative leaders, and chief state school officers. When NAEP was first created, ECS was granted the original contract and administered NAEP for 14 years before the bid was won by the Educational Testing Service. ECS sponsors an

annual assessment conference in June that brings together researchers, policymakers, and state assessment practitioners.

- ***State Consortium on Testing***—The Chiefs' State Assessment Center is also involved with a consortium of more than a dozen states (the number keeps growing) interested in developing performance-based assessments. In a three-way partnership, the Chiefs are identifying states' technical needs; the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA is conducting the research on types of instruments, and the Education Commission of the States is communicating with policy audiences on the issues related to performance assessment.
- ***President's Education Policy Advisory Committee***—This committee of representatives of business, education and the media is charged with providing ongoing advice to the President on a variety of education issues. Some members have called for a national test of every pupil to measure progress in student achievement at the 4th, 8th, and 12th grades (the same grades as specified in the National Goals). Their recommendations were due in 1991.
- ***Commission on the Skills of the American Workforce***—Based at the National Center on Education and the Economy, the commission has proposed a series of national performance exams that students would complete by the age of 16. The results could be used by employers as a job credential.
- ***Secretary's Commission on Achieving Necessary Skills (SCANS)***—Appointed by the Secretary of Labor, SCANS has convened task forces grouped by industry to determine what general functional skills students need to be ready for work. The commission will identify skills and methods of assessing them. Reports were expected in May, 1991, and February, 1992.
- ***Congress***—The National Report Card Act of 1990, introduced by Sen. Jeff Bingaman (D-N.M.), would have created a report card to measure educational achievement of students and schools and establish a National Council on Educational Goals to determine what indicators would be used. The bill, which did not pass, is expected to be reintroduced in 1991. The Congressional Office of Technology Assessment conducted a comprehensive study of alternative assessments, uses of assessments, and uses of technology in testing due for completion in fall, 1991. The General Accounting Office may also study assessment for Congress. Such studies are usually a prelude to legislation.
- ***U.S. Secretary of Education***—Has issued annually a much-derided "Wall Chart" that compares states on per-pupil expenditures, pupil-teacher ratio and other "input" variables. The only information on student performance is average SAT and ACT scores, which has raised the ire of many educators. This ritual will be discontinued, new Education Secretary Lamar Alexander announced after taking office.
- ***Educate America***—This most recent entry into the national test game is a nonprofit organization chaired by former New Jersey Governor Thomas H. Kean that called early in 1991 for high school seniors to take a national examination of their knowledge and skills. This group, which is seeking help from Congress to fund the test, proposes that all take nine hours of tests in reading, writing, mathematics, history, geography, and science. Results would not affect graduation, but scores would be sent to prospective employers and colleges.
- ***FairTest***—The National Center for Fair & Open Testing based in Cambridge, Massachusetts, issued a statement signed by 76 education and civil rights organizations and individuals who are opposed to expansion of the National Assessment of Educational Progress and the concept of a single, national test to measure progress toward national goals. This consortium has adopted position papers on the role of assessment—and the role of the federal government—in education, and is plotting a course of action to fight the adoption of such proposals.

The debate begins

Perhaps it was the final recommendation in the NAEP's 20-year report—"to join in a concerted effort to achieve our national goal of improved academic achievement for all students"—that prodded some of the new players in Washington to get into the game. For 20 years, they suggested, the states have been trying on their own to improve the academic performance of their students, but it hasn't worked. Perhaps it's time for a new approach.

All of this activity on the national level—and its implications for the school and classroom—made many educators nervous. As noted above, most of the nation's educational leadership organizations joined with FairTest to protest the movement and take an active role in lobbying against it. But as the AASA survey on assessment indicated, the grass roots members don't find the idea especially threatening—until they get around to comparing school districts and schools.

Grass roots reaction

Respondents to the AASA survey gave a qualified vote of approval to the "new direction being taken by NAEP in

establishing student objectives as well as measuring progress." A surprising 65 percent called this a "welcome resource," although a number of those qualified their answers by penciling in words like "maybe" or "it remains to be seen." The remaining 35 percent checked the response "A dangerous trend."

The written comments to this question show that most of the respondents were willing to give the NAEP the benefit of the doubt:

"Let's do it. It's a *real* world. I'm 100 percent for accountability. It's the only way things will change "—David Romstad, superintendent, Parkview, Wisconsin.

"A national core curriculum/assessment would provide some consistency in basic skills taught, by grade, which would benefit student transfers"—Delores Quinn, testing supervisor, McComb, Mississippi.

"NAEP, when all is said and done, is another ETS bureaucracy. Currently, I am dubious about their results—or interpretations—based upon multiple choice tests. Therefore, I welcome this new direction, I think?"—Dick Lundquist, guidance director, Westside Community Schools, Omaha, Nebraska.

The second part of the question dealt with suggestions that NAEP release test results state-by-state, district-by-district and school-by-school. This part found the respondents decidedly more negative: over 56 percent called state-by-state comparisons "a dangerous trend," a figure that rose to 71 percent each for "district-by-district" and "school-by-school."

Once again, the written comments ran the gamut:

"This appears to be a current trend and one which we will have to live with for accountability reasons, whether justifiable or not."—Judy B. Howard, assistant superintendent, Crawford County, Ohio.

"Comparisons are always dangerous, but I fear we've brought it on ourselves." —Michael Pomarico, superintendent, Atchison, Kansas.

"A two-edged sword as all testing. May have simultaneous benefits and liabilities."—Alex Epanchin, testing director for the Pinellas County, Florida, Schools.

"I am ambivalent. I am a little suspicious. Will this be a new "Wall Chart? Will test results become our ultimate 'outcome?'"—Lundquist.

"The closer it comes to home the more uncomfortable I become. When test scores are used by themselves without comparative socioeconomic data, etc., more harm than good can come of it."—Gary Hartman, testing specialist, Banks, Oregon.

"National state-by-state comparisons may encourage more federal and state funding to increase student achievement. Such listing may be 'healthy competition.' However, other listings would be demoralizing and counterproductive, causing an elitism that would be detrimental to poorer districts and schools."—William D. Bartman, test coordinator, Pottstown, Pennsylvania.

The pros and the cons

There are pros as well as cons to the question of giving a single national test to measure progress toward national goals. Here are some of the key issues on either side along with comments by some of the players actually participating in the fray or viewing the action from front-row seats:

Local control

Many of the critics fear that a national test would exert too much influence over local curriculum and instructional practice. For example, they speculate, if all students are tested for graduation, and teachers' evaluations are based on how their students perform, there will be incredible pressure to teach to the test. But such standards imposed from a national level run counter to the research on school improvement that says change should emanate from the school and classroom level, where instruction can be tailored to the needs of the students.

Several experts queried for this report see no conflict between setting general national standards and allowing local schools the flexibility to choose the best way to achieve those standards. As an example, NAGB contracted with the Chiefs to develop a consensus among the states on the general framework of what should be measured in the NAEP math, reading and science assessments. While the statements are specific to grade levels and subject areas, they are general enough to permit flexibility at the district and classroom level.

Ramsay Selden, director of the assessment center at CCSSO, provided this example: In fourth-grade reading, students should be able to accomplish the following tasks with regard to reading prose literature, informational materials and documents:

- Form an initial understanding and create a coherent representation in the student's own mind of what was read;
- Extend comprehension with details and elaborate on that understanding by organizing and integrating the information;
- Develop an evaluative judgment of the passage, such as whether the argument was valid or the prose beautiful, and state the basis for the student's judgment.

While adhering to these general frameworks, the local curriculum could use either phonics or whole language to teach reading and make local decisions whether to use textbooks or trade books and which ones to use.

"To the extent there is a body of knowledge that everyone agrees that students should know, such as 2 + 2 = 4, NAEP and other tests *should* have an impact on accountability and curriculum," contended Richard Boyd, new chair of the National Assessment Governing Board.

Standards: Good or bad?

The move to set national standards of achievement is unprecedented, and some critics fear it could produce undesirable effects. The 1990 NAEP math assessment released in June 1991 for the first time reported results according to the proportion of students who achieved at basic, proficient, and advanced levels. The National Assessment Governing Board convened a large group of judges, including math teachers and district specialists, a PTA official, school board members and college professors, to rate test items at one of those levels. Participants' comments on the standard-setting ran the gamut:

- This process is usually done by a small group to set a single standard instead of three standards for the whole country.
- It was difficult to distinguish between items to be rated "proficient" and "advanced."
- When the assessment reveals large gaps between what students *should* know and what they *do* know, more parents and citizens will be involved in improving education.
- Results will give more ammunition to those who would bash schools. Educators will try to ignore the results or find reasons why they don't apply to them.

In the past, NAEP reports received "little attention," commented NAGB director Roy Truby, "because no one ever said what performance was good enough." Instead, the reports would say things like, "Sadly, only 60 percent of 17-year-olds could write an acceptable letter asking for a job."

At the press conference releasing the writing assessment, reporters asked, "Wouldn't you expect 100 percent of the seniors to be able to write a letter?"

"Not necessarily, Truby responded, "because they only had 15 minutes to do it."

In the past, NAEP reported trends over time, such as how American students were performing compared with 20 years ago. "Most people today are more concerned about how we're doing compared with Germany, Japan

Two Decades of Reporting Educational Progress

The Nation's Report Card

The National Assessment of Educational Progress (NAEP) is a federally funded program to measure, over time, what American students know and can do in various subject areas. Also known as the Nation's Report Card, NAEP has, since 1969, conducted assessments in reading, mathematics, science, writing, history, geography, civics, and other fields. The assessments are given to a sample of students ages 9, 13, and 17. Some assessments are also made of young out-of-school adults in their early 20s.

The Report Card is a project of the National Center for Education Statistics in the U.S. Department of Education. Contracts are awarded to qualified organizations to conduct the assessments. For the first 15 years, NAEP was operated by the Education Commission of the States in Denver, Colorado. The current contractor is the Education Testing Service in Princeton, New Jersey.

When NAEP was first created, it was specifically prohibited from releasing data on a state level. Instead, student achievement was reported for the nation by age and by region. Results were further broken down by age, gender and ethnic group. In 1988, Congress approved the National Assessment of Educational Progress Improvement Act, which changed NAEP in two significant ways:

- A National Assessment Governing Board was established.
- NAEP was authorized to conduct voluntary, pilot state assessments in math and reading.

Sets policy

The National Assessment Governing Board sets general policy for the assessments, including:

- Achievement goals for age and grade.
- Assessment objectives, test specifications, and methods.
- Guidelines and standards for analysis of data and reporting and disseminating results.
- Ensuring that test items are free from racial, cultural, regional, or gender bias.

The act specified that the Governing Board should contain a cross section of elected officials, educators and the general public, including governors; state legislators; local and state school board members; classroom teachers; testing and curriculum specialists; parents; and representatives of the business community.

In one of its first major activities, the board voted to set standards for reporting achievement according to the proportion of students performing at "basic," "proficient," and "advanced" levels.

The Nation's Report Card was scheduled to release the results of its 1990 math assessment, by state and by achievement levels, in the fall of 1991.

and other countries than they are with how we did in the 1970's," Truby explained.

It was for these reasons that the governing board began a process to determine criterion-referenced standards and one of the key issues was whether to set a single standard or multiple ones. There was some opposition at the start to multiple levels because it suggested academic tracking. But at public hearings, Truby reported, "The chiefs, governors and others said that such distribution exists and looking at the dynamics of the distribution provides useful information for setting goals."

If standards had been in place for the last four or five assessments, and you look at basic, proficient and advanced as three floors of a house, "What's happened is that people are abandoning the top and bottom floors and moving toward the middle," Truby said. "Students in the top 15 percent are dropping off, and most people don't even realize it."

Reporting such levels will allow policymakers to address such shifts, which are invisible when only averages are reported, he concluded.

Time for testing

One of the "cons" is the contention that there is already too much testing, which takes away from instructional time. More testing, they say, would reduce instructional time that much more. "We've tested our brains out for 10 years and nothing's gotten better," said D. Monty Neill, associate director of FairTest. When pressure comes to raise test scores, teachers spend time teaching to the test. "I've even heard people from (commercial) testing companies say that education should not be focused on raising test scores," Neill added.

While Boyd agreed that "tests are being piled on top of one another," he added that the national goals have created a new reason for testing: to satisfy calls for accountability.

He pointed out that NAEP tests only a *sample* of students, about 2,000 in each state. This would not create a big impact on the classroom, Boyd said. But to minimize any further impact on classroom time, Boyd suggested looking at some of the commercial tests being used in schools with an eye toward eliminating them. "I'm not sure all the standardized tests are that necessary," Boyd commented, "particularly in states that have developed criterion-referenced tests."

Test format

Critics argue that national tests are overly reliant on multiple choice items while students should be spending more time on problem-solving. "There's too much drill and kill instruction going on, and more testing may promote that," Neill said.

But Truby pointed out that NAEP used open-ended questions for the first time on its 1990 math assessment. These tests also included more analytical questions and permitted students to use calculators for the first time.

"If you look critically at the open-ended questions, most required short answers," said Truby. "Is the math field completely happy with that?" No. Will they ever be? Probably not. But that was the consensus in 1988 and at that time, it was the state of the art."

"Everyone would like to have more sophisticated items on the tests," Boyd said. "Many governors understand the tests are not what they ought to be. But do you say 'we're not going to hold the schools accountable for the next 10 to 15 years until tests get to where they need to be?' Politically, that just isn't going to fly."

As for alternatives, performance-based assessment is "frightfully expensive," Boyd added, noting that oral reports and portfolios cannot be machine-scored. "There's a real question about whether the resources will ever be allocated to do performance assessment, at least on a large-scale basis."

On the contrary, countered Selden, more than half the states are already doing some type of writing sample, using teachers as scorers. If the reading and writing were integrated with a social studies or science assessment, the states would get even more bang for their buck. The way to make performance assessments affordable, Selden concluded, is to: (1) simplify the task, (2) use "cheap" materials, and (3) involve the classroom teacher.

"Scoreboard mentality"

When state-by-state comparisons become available in the 1990 NAEP math assessment, some people are concerned that this will lead to a "scoreboard mentality" that will rank states without giving meaningful information that can be used for improving schools.

"Nonsense," responded Truby, who is himself a former chief state school officer. The math assessment will reports results by subject area, including algebra, operations and measurement.

"If I were still superintendent in West Virginia, I might look at the math assessment for Kentucky, whose scores are usually similar." If he found that the eighth grade algebra scores in Kentucky were much higher, Truby said he might look at the kind of textbooks Kentucky uses, the preparation required of the teachers, and other factors where the two states were different.

"There's a wealth of information," he exclaimed. "If you don't see more than a box score, it's because you haven't paid attention."

"Do it right!"

In the consensus process for the NAEP state-by-state assessments, the Chiefs solicited comments from the states as

The National "Achievement" Goal

The historic Education Summit in Charlottesville, Virginia, was convened in September 1989, in an atmosphere of growing urgency among political and business leaders about the success of American education.

The President and the nation's governors adopted a set of national education goals, which were highlighted in President Bush's State of the Union address and formally adopted by the President and governors the following February.

The six goals focus on (1) ensuring that children start school ready to learn, (2) raising high-school graduation rates to 90 percent, (3) increasing levels of educational achievement, (4) promoting science and mathematics achievement (5) as well as literacy and lifelong learning, and (6) freeing schools of drugs and violence.

The goal targeted to student achievement, which is most relevant to this report, and its accompanying objectives, are as follows:

National Goal:

By the year 2000, American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy.

Objectives:

The academic performance of elementary and secondary students will increase significantly in every quartile, and the distribution of minority students in each level will more closely reflect the student population as a whole.

The percentage of students who demonstrate the ability to reason, solve problems, apply knowledge, and write and communicate effectively will increase substantially.

All students will be involved in activities that promote and demonstrate good citizenship, community service, and personal responsibility.

The percentage of students who are competent in more than one language will substantially increase.

All students will be knowledgeable about the diverse cultural heritage of this nation and about the world community.

to whether the assessment should measure what is being taught now or what students *ought* to be learning. This includes the question of whether to demand more intellectual rigor and higher-order thinking skills: If those skills are not measured, how will policy makers know if they are being taught?

"With one dissenting vote," Selden said, state chiefs and their curriculum managers "clearly felt we should be driving things forward somewhat with the basis of these comparisons." The message was, "Do it right!"— frame the statements in terms of what students should be able to do, even if only 5 percent of the students are now reaching those objectives. "If we want that skill level across the board, so be it," he said.

Reaching consensus

Obtaining agreement on what all students should know: There's the rub. Some critics argue that the testing "cart" is getting ahead of the national goals "horse." The president and governors have agreed to some broad goals and objectives, but little discussion has occurred on how those goals will be implemented.

To further complicate the issue, several groups tackling the question are reaching parallel, but slightly different, conclusions about the need for a national yardstick to ensure that students are achieving high levels of skill and knowledge.

For example, a group with strong backing of the business community—the Commission on the American Workforce was funded by the Carnegie Corporation—is leaning toward performance-based measures that would be administered to all students at the local level. Some voices are still pressing for a single national exam, but as the discussions extended into 1991, there seemed to be more interest in developing a flexible system that sets national standards but gives states and districts leeway in how those standards will be met and measured.

As this report goes to press, most groups are still deliberating, so predictions are hazardous. Interviews with representatives of these groups, however, point to three possible directions:

1. Testing a national sample of students, through NAEP, to determine progress on the goal of raising student achievement in subject areas.
2. Establishing performance-based demonstrations of competency that districts could use as an exit exam. Students could make presentations, produce portfolios and carry out other hands-on activities to demonstrate proficiency for graduation.
3. Developing a series of tests or demonstrations that would be available on a voluntary basis and would be "calibrated" to be equivalent to NAEP items.

What lies ahead?

The assessment game has been played on the national scene for about 20 years, but until recently in an empty stadium.

Now, with a new emphasis on goals, fan interest is increasing dramatically. But educators in the back row are concerned because there seem to be three or four teams on the field at the same time.

It appears that the game will keep evolving for the next few years, and local districts may find it confusing to play because the rules will probably change more than once.

But they might be able to benefit from the debate, for example, by reflecting on the degree to which standardized tests administered in their own district may suffer from the same flaws attributed to a national goals exam. If commercial standardized tests also rely on multiple-choice tests of low-level skills, would districts be better off discarding them too, as some experts have suggested?

Most important, however, many states and districts are not waiting for a federal mandate. At least two multi-state consortiums are looking seriously at performance-based assessment—indicative of the efforts under way to tackle the assessment problem on their own rather than have answers prescribed from above.

Preparing Students for the World of Work

Is Another 'National Test' on the Horizon?

The Commission on Skills of the American Workforce, based at the National Center on Education and the Economy NCEE), issued a report in 1990 that called for far-reaching changes in the way students make the transition from high school to work and higher education. The report provoked a lot of discussion and raised the specter of a national test and national standards—one that at least for then did not involve NAEP.

The report, *America's Choice: High Skills or Low Wages!* proposed that all students should, by age 16, be required to pass a series of "performance-based assessments" that were based on national standards. Students would be awarded a "Certificate of Initial Mastery" that would permit them to go to work, enter a college preparatory program, or study for Technical and Professional Certificates.

The assessments would be "new performance-based examinations for which students can explicitly prepare," the report said. Characteristics of the system are that:

- Students would have multiple opportunities for success.
- The exam would not be used as a sorting mechanism as is the case with most national tests.
- The goal is to set a tough standard that almost everyone will reach, although not all at the same time.

As a follow-up to the commission report, (NCEE) and the Learning Research and Development Center (LRDC) at the University of Pittsburgh have forces to develop a prototype system of standards and exams that could be used by states and local school districts.

More than just tests

The task is larger than designing tests, cautioned Lauren Resnick, director of LRDC, who has frequently written about the dangers of tests driving curriculum. "We don't think establishing a test or several tests is what the nation needs," Resnick explained. "We need a system that, over time, will produce the kinds of assessment and the kinds of instruction that are needed."

The two centers want about a dozen states and six or seven school districts to participate in the 18-month pilot program. The senior education officials in the states and cities would provide overall direction to the process. Teachers and other specialists in participating school districts and states will be heavily involved in the design and field testing of the assessments.

The goal is to set standards and design measurements that will be reflected in the curriculum, teaching methods, and staff development of the participating districts. The states and school districts being considered are involved either in performance-based assessments or school restructuring efforts.

These new assessments are intended to measure high levels of knowledge and performance, including applications of skills and knowledge to real world problems. Student competence could be demonstrated in three ways:

- Performance examinations given at particular points in a student's career.
- Projects that allow students to demonstrate their ability to solve complex problems over time, often working with other students.
- Portfolios assembled from work students do over months or years.

The subjects to be assessed and the grade levels will be determined by the participants.

CHAPTER NINE

The Power of Communication

To some, the word "communication" might only signify releasing test scores. But if used to its fullest extent, it can bring understanding and support from staff and community alike.

Most veteran educators can remember scarcely two decades ago when a school's test scores were as big a secret as the combination to the safe. And how appropriate, because that was probably where the scores were kept, far from the prying eyes of parents and the local press.

It literally took an act of Congress to bring test scores out into the sunshine, and before very long it was obvious why they had been hidden all those years. Few people really understood what they meant, and when newspapers began ranking local schools from "best" to "worst" based on standardized test scores, the rules of the game were changed forever.

The lesson to be learned was that the public, as well as the federal government, demanded accountability for the money spent on education. Furthermore, thanks to the media's post-Watergate mentality, the harder school systems fought to keep them under wraps, the bigger the headline that proclaimed, "Local Scores Plummet."

Protestations were futile. Editors resented being told that they did not understand what the tests meant—that a two-percentile decline was not a "plummet" or that they couldn't call a school with a 68 average "mediocre" because one down the road averaged 70.

It was a difficult period for school administrators, many of whom stubbornly resisted the inevitable. As late as 1974, a survey by the Educational Research Service showed that 68.3 percent of responding districts did *not* release their test scores. Stated reasons ranged from the obvious—"The public and press might misinterpret the test results since few understand the subtleties of standardized testing" to the desperate—"Releasing standardized test scores violates the First and Fourteenth Amendments and denies children 'their right not to be stigmatized by degrading titles.'"

The battle, however, was a losing one, because if school districts did not release their scores willingly, somebody else would do it for them. If that somebody was the state, it invited the inevitable ranking. In those cases where citizen activists went to court to force the release, the outcome was assured because of the federal Freedom of Information Act or new state sunshine laws. The ensuing publicity threatened to cost many a career educator his or her job.

At the time, virtually the only means available to prove "accountability" were the standardized norm-referenced achievement tests customarily used for monitoring program effectiveness. Most educators agreed that they were not really appropriate for comparing the performance of schools or school districts, but what else was there? Did schools really expect the public to take their word for it while they slipped the scores back into the safe?

Even today, most of the test scores being reported in the name of accountability are still not appropriate, but as we have seen, many promising movements are under way to develop assessments that are.

And just as important, the development of strategies for telling the story of student performance has begun to change the adversarial atmosphere into one of cooperation.

A never-ending process

Just as assessment is a continuing program aimed at improving instruction, so is communication a never-ending process. It is always appropriate to show the public how schools are succeeding in improving instruction. Problems can arise when standardized test scores are released out of a communication vacuum because a context is lacking. If the scores seem high, the public will often accept the report with little reaction. But if they are low—or if the local media seize upon them to scrutinize school effectiveness, trying times may lie ahead.

It would be ideal if there never had to be an annual release of test scores per se—if the schools could issue periodic progress reports in a format other than columns of numbers. Unfortunately, however, either because of state mandates or their own school board policies, most districts must administer standardized tests, and they do not have the luxury of withholding the results. The word "accountability" has become a part of the education vernacular, and

while there is much work under way to define "accountability" as more than test scores, in most cases those columns of numbers are still the focus of the report.

For those schools that must participate in such an annual ritual, the key to public understanding is to provide a living context for the assessment results. If the key members of the community—parents, taxpayers, school board members and politicians—are aware of the continuing progress in improving student achievement, they will not react inappropriately to test results.

Therefore, long before it becomes time to make public the results of an assessment program, the groundwork should already have been laid based upon the following considerations:

Involvement fosters understanding

It has become virtually a truism in American education, verified year after year by Gallup polls, that people who visit school buildings are far more likely to be supportive of the local educational system than those who do not. It stands to reason, therefore, that the first step in building understanding of the school program is to bring citizens inside the school buildings, preferably to involve them in the educational programs. To them, the standardized tests become only one of many measures of student performance.

The use of parents as volunteers has been a boon to schools in providing additional support to teachers at little or no extra cost. While it may be possible to quantify the contributions these parents make in preparing materials, reading to children, grading papers, and so forth, the impact of the volunteer program on fostering understanding of the school program is immeasurable.

Those parents who work with the classroom teachers are part of the continuing process of evaluating student progress. They are not likely to be swayed from their support of the school by columns of numbers, because they see for themselves what the children do every day in class. By the same token, therefore, bringing other members of the community into the school, by whatever pretext, can have a similar result.

The rapid growth of school-business partnerships, for example, has brought a new dimension into the classroom through increased resources and materials, and the involvement of businesspeople (many of whom are not parents of school children) in day-to-day instruction. Just as with parent volunteers, these "partners" develop a feeling of ownership in the school and would resist unwarranted attacks on the educational system based on a misinterpretation of test scores.

Similarly, the widespread involvement of senior citizens as volunteers in the classroom has broadened support from that growing segment of society. Senior citizens may no longer have a direct relationship with the schools, but they do vote, and they do have influence in many institutions throughout the community.

The media as partners?

It is an unfortunate fact of life that in many communities the schools and the press maintain an adversarial relationship. Some journalists may feel that their role as the public "watchdog" requires a cynical attitude about government functions. Since the schools spend public money, they are considered fair game. On the other hand, some educators are just as suspicious of the local media. They are afraid that all they want are sensational headlines and are hesitant to open up to them.

In very few cases these attitudes are justified, but the lack of mutual understanding of each other's roles usually exacerbates the situation. Thus when an "event" happens—in this case the release of standardized test results—the fallout can be very damaging to the schools.

It is a mistake for educators to perceive the local media as their natural enemy. The people who run the local newspaper may be products of the local schools or members of a PTA. As citizens they are vitally interested in student achievement, and as professionals they have much to gain as part of a community that feels good about its schools. A sensational story "sells papers" for a few days. A good educational system sells the community forever.

It is, therefore, to everybody's benefit to foster a continuing relationship between school and media far removed from news events and press releases. It stands to reason that if local reporters, editors, news directors, and so forth, were frequent visitors to the school—whether to cover stories, speak to students or just have lunch with the principal—they would still report the news but they would be less likely to allow a single event to color their judgment.

The journalist's role

The strategy is to make the release of test results one of many "events" that demonstrate the progress of students. And journalists should welcome the opportunity to cover them. Lisa Walker, executive director of the Education Writers Association, says that her organization's workshops stress the importance of the local media going behind test scores to show the "true picture" in the schools.

"We don't believe in just reporting test scores," she said.

"We want to find the real measures of success—things like drop-out rates, improvements over previous years, and whether the schools are meeting the parents' expectations."

Walker conceded that many education writers are criticized for having "the usual knee-jerk reaction" to the release of test scores and she did not try to defend them. However, she raised this note of warning: In many areas reporters' research raises the "concern that testing is driving the curriculum," and thus they assume the responsibility to raise the issue.

In the last decade, Walker pointed out, "The education beat has gotten more professional, and a lot of the criticisms of the press may be unfair." Her advice: Treat these reporters as the professionals they are, and their coverage will reflect it.

On small newspapers, reporters usually have several beats and cannot be expected to understand the subtleties of education, she said. But those are usually in small towns where educators and journalists can develop personal relationships that can have the same effect.

The teacher's role

It was evident from the responses to the AASA survey that a vast majority of school districts do not brief teachers about the meaning of the standardized assessment program. An hour inservice in test administration is not the same thing as ensuring that teachers understand how these tests fit into their overall instructional program and what the results might mean to the image of the school and the school district.

We have already discussed the potential abuses if, once a year without any preparation, teachers are handed a package of standardized tests and are led to believe—whether by statement or inference—that high scores are a district priority. Thankfully, incidences of cheating are rare, but the above scenario fails in two important ways:

- The teachers are not helped to understand how these standardized tests fit in with their everyday classroom assessments and how they can be used to strengthen instruction.
- The teachers are not enlisted in the school system's effort to communicate a true picture of student performance.

Building-level inservice

"It's important to recognize that the testing courses that teachers take at the university relate not one bit to real life," said Carole C. Gran, principal of James Lewis Elementary School in Blue Springs, Missouri. Thus, she added, "They are very uncomfortable with the statistical side of the program and may reject it as not being connected to teaching."

Gran feels it is her role as principal to "eliminate the misinformation" and she does it with "a lot of inservice at the building level on how to interpret and understand tests."

Since Missouri standardized test results are not available until after the school year ends, Gran holds her sessions the following August. She shows her staff the school data, grade data, and trends so they can analyze all of it, and then they discuss what it all means in planning for the upcoming school year.

In this way, she feels, the teachers gain an understanding of the testing program and how it fits into daily instruction. And, of course, when parents want to know what all the numbers mean, the teachers are able to tell them.

The timing of such sessions will depend on the timing of the tests, but the cycle must begin long before the tests are actually administered. In addition to their value in refining instruction, such sessions demonstrate that in the context of the ongoing communication program, the teachers should be first a prime *audience* so they can later become key *communicators.*

To each his/her own

Enlisting teachers as enthusiastic members of the assessment team was one of the major goals of the California Assessment Program as it developed the statewide performance-based Writing Achievement Program in grades 8 and 12. "The buying-in had to be at the teacher level," explained CAP specialist Barbara Weiss, "because we needed 700 of them to grade the papers in addition to those who helped develop the tests." Although the teachers received an honorarium, they still had to be willing to work on weekends and holidays.

Giving each teacher his or her own copy of the special "Report to Teachers" for each school was one of the keys to maintaining their support. "Originally we sent them directly to the teachers because we felt it was so important," Weiss said, but there were repercussions from principals and superintendents. The next year there were separate reports for administrators, but teachers still got their own.

In addition to explaining the eight sections of the Writing Achievement tests and the grading philosophy and offering special tips on using the results in the classroom, the reports also offered comments from teachers who scored the essays. These quotes show their enthusiasm for the program and how their increased knowledge of evaluation has improved their teaching styles and their students' performance.

"Those reports were probably as much public relations as anything else," Weiss said, but it worked. A survey of teachers by a University of San Francisco professor re-

ported both the acceptance level of the program and the perceived improvement in teaching as well over 90 percent.

What about the board?

Probably nobody is more interested in assessment results than the members of the school board, but are they any more apt to understand them than the parents are? Board members are as susceptible to hasty judgments as the average citizen unless they too are involved in the process long before the report is dropped in front of them at the board table.

The district's assessment program should be made a major agenda item for the board long before scores become an issue. Staff members, including teachers and principals, should be on hand to discuss how the assessments are used to improve instruction, what other factors are considered in making judgments about school performance, and what steps have been taken as a result of last year's results.

While such sessions are especially valuable in the case of newly elected or appointed board members, even the veterans will find the information valuable. This will also serve to focus their attention in appropriate directions when the test report itself comes up for discussion.

Recruiting Parents as Partners

When a school sends a note home to parents about their children's test scores, they can be assured of pretty good readership—perhaps the best of any home-school communication all year. Don't fail to take advantage of it.

Whether the test scores that accompany the message are good or bad, the message will find the parents in a receptive mood as it reminds them of their vital role in the educational process.

Provide a little checklist for parents, like this one with items culled from a number of school districts and the National PTA.

Parents: No matter how your children scored on these tests, you have an important role to play in our continuing effort to help them realize their fullest potential. Here's what you can do as a partner in your children's education:

- Become more involved. Parent involvement helps students learn and improves schools.
- Provide a comfortable and quiet place for children to study and do their homework. Stock it with good books and reference materials.
- Ask your children about school things. Demand to know about grades, test scores, etc. Look over their homework to show that you are interested in what they are doing.
- Set a good example. Parents should show their children they believe reading is enjoyable and important. Parents shouldn't spend all their time in front of the TV either.
- Become involved in school activities, and not just in the elementary grades. It shows your children how much you value school—and helps their teachers do their jobs better.
- Encourage your children to do their best in school. Don't ever let them forget how important education is to you.
- Emphasize academics throughout your children's school career. Don't get too caught up in athletics or in preparing your child for the world of work when academics should be your prime concern.
- Support school rules, discipline and goals.
- Encourage your children to do their best, but don't apply too much pressure by setting unrealistic goals or involving them in too many activities.
- Call teachers early if there is a problem. Go see them, or the counselor or principal, while there is still plenty of time to rectify the situation.

While these tips are valuable throughout the school year, you can add three more to highlight just for the testing season:

- Encourage your children to be serious about the tests and not see them as a nuisance or a break from "real" school work.
- Try not to upset the child before the tests, such as by introducing family conflicts, etc.
- Make sure the child gets a good night's sleep before the tests and eats a good breakfast that morning.

Public Relations: The name of the game

All of the steps discussed above in the ongoing communication process can be described in two words: public relations. It is practiced to some degree in every school district, under many names: school-community relations, public information, communications. In some cases it is not practiced very well, perhaps because of a misunderstanding of its role.

Several decades ago, the American Association of School Administrators defined school public relations as a "cooperative development and maintenance of efficient two-way channels of information and understanding between the school, its personnel, and the community."

Twenty years after that, the National School Public Relations Association (NSPRA) substituted the word "educational" for "school" and defined the process this way:

> "Educational public relations is a planned and systematic two-way process of communication between an organization and its internal and external publics. Its program serves to stimulate a better understanding of the role, objectives, and accomplishments of the organization. Educational public relations is a management function that interprets public attitudes, identifies the policies and procedures of an individual organization with the public interest and executes a program of action to encourage public involvement and to earn public understanding and acceptance."

While newer textbooks further refined the definition of public relations by adding such words as "dynamic" and stressing "involvement" as an integral part of the process, the key words continued to be "planned," "systematic," and "two-way."

What this means is that the act of telling the school story—for example, the annual assessment report—should actually be a year-long process and that merely "informing" is not enough.

In the context of public relations, communication is only one part of the process and doesn't even come first. In fact, if the first step is to publish a newsletter or hold a press conference, the campaign may come up short.

The four-step process

If there were such a thing as a generic assessment report, we would feature it here as a model to follow. As the previous chapters have shown, however, there is an almost endless variety of assessment programs in use around the nation, each with unique complicating factors. But a great majority of school districts are using some kind of standardized measures in their assessment programs. And whether they are norm-referenced or performance-based, there are always numbers in the report.

The following discussion should be general enough to apply in virtually all situations where a school or school district is conducting an assessment program and is about to tell the world about it.

That challenge is met by invoking the four-step process that is the heart of all successful public relations programs:

1. Analyzing
2. Planning
3. Communicating
4. Evaluating.

Step 1:

Analyzing

If the goal were simply to release test scores, it would be simple enough. But that assumes the tests were administered for the sake of testing and that the scores can stand on their own. But that is not true. The assessment program was designed to measure and ultimately improve student achievement. It did not begin after the tests were scored, and neither should the communications program.

The "hook," as they call it in the media, consists of the test results; but the purpose in publicizing them is to show the community how assessment is related to improving school performance.

All too often people involved in publicizing test results think of the public as one faceless monolith. This is a serious mistake. To a skilled communicator, there is no such thing as "the public." There are many publics out there, each with its own unique background, needs and degree of involvement. And since the assessment program would affect many different publics in different ways, each may require a unique communication effort.

The analysis phase of *any* public relations effort—not just releasing test results—begins by thinking about all of the publics that might warrant consideration. Here is a list a school superintendent might come up with as the potential audiences of his or her assessment report, not necessarily in priority order:

- The school board
- Principals
- Teachers
- Nonteaching school employees
- Parents
- Nonparent taxpayers

- The press
- Television and radio reporters
- Statewide curriculum groups
- Educational associations and unions
- College professors/researchers
- State department of education
- Legislators
- City/county officials/councils
- Citizen groups
- Parent-teacher associations
- Advisory councils
- Realtors
- Local businesses
- Other governmental agencies

Note that a "public" can range in size from one person to tens of thousands, and each can be very important to the school and the district. The challenge is to select those "publics" that need to be addressed directly on this particular topic, and determine what message is correct for each and when it should be delivered. This involves asking several questions about the people in each public:

- What do they already know about the subject?
- How much about the subject do they want or need to know?
- What are their specific concerns?
- What is the anticipated outcome after they receive the message?

It's never easy

A school system does not have to be at the low end of the spectrum to face a major public relations challenge. Take the suburban district of Englewood Cliffs, New Jersey, where expectations are so high, said Superintendent Gregg R. Hauser, "that when our kids averaged 95.5 in math, the school board demanded to know why it wasn't 96."

What Hauser faced in 1991 was a new version of the test, and new norms. "We figure that the scores will be 5 to 10 percent lower than last year," he said, and in his case even a 90th percentile average could mean trouble. By anticipating the reaction—by the parents, the school board and the community—Hauser was able to start early in the school year, months before the tests were administered, to prepare the public for the new numbers, primarily through parent meetings and brochures. More meetings were to be held after the tests were given "in case they forgot."

That type of misunderstanding, as pointed out earlier, is one of the main reasons for the movement away from nationally-normed standardized tests. But in New Jersey, Hauser said, he has no choice because of state mandates, so he must adapt.

"An education reporter from an urban daily newspaper went to work for the city's school district as an information specialist. After only a couple of months on the job, he called his former boss—an editor to whom he had been close—to complain, gently, of the negative coverage the paper had been giving the school district. "Listen," the editor said, "the people you work for are educators. They're supposed to be educating kids. They're paid by the taxpayers to educate kids. When they do, that's not news. That's what they're paid to do. It's when they don't that it's news."

—Building Public Confidence for Your Schools, NSPRA.

Analyzing the publics

While it would be impossible to develop a hypothetical scenario that would cover all aspects of all audiences, here are some examples of how this kind of analysis can determine the character and the timing of the communication:

The school board

The information is critically important to them and their status in the community. Because of past informational programs, they are aware of the ongoing program and need the unvarnished results plus explanations of anything that might lead to questions. The report, as issued by the testing authority, should go directly to the board along with explanations by district officials. Obviously, board members should not read about it first in the local newspapers, so the information should be delivered prior to public release if possible. Since the board's reaction affects future programs as well as job security, the results should be accompanied by planned follow-up activities, especially where scores are disappointing.

The principals

Principals often feel their necks are on the block, so to speak, so they should be informed as soon as the information is available, probably before anybody else. This will enable them to conduct meetings with their school staffs to inform them about the results, and explain the implications, before they read about them in the newspapers. The superintendent can tell the principals about the district picture and what stance the administration will be taking. The principals should be asked to advise the superinten-

dent on potential follow-up activities for the school board report and subsequent communications. Because of the sensitive nature of this interchange and the need for two-way communication, this should done face-to-face, with materials distributed in person.

The teachers

Except for the students, they are the most personally involved in the results, so the presentation should be made personally by the superintendent, principal, or other appropriate administrator just before the public release. The teachers are familiar with the testing program; they need to be brought up to date on the results and their implications for the school and the district. The teachers' immediate role at this point would be to answer parents' questions and concerns, so time should be allocated to discussing strategy. Eventually, of course, all the teachers may be expected to respond to the test results with modifications in instruction, so that should be touched on as well. Later meetings will be devoted to how the results might directly impact on instruction.

Nonteaching employees

These staff members are often left out of this process. While it is true many of them often don't see a need to understand fully the educational implications of the program, they are important public relations agents in their communities, and they should understand the meaning of the scores and what the schools intend to do as a result. Ideally, they should be part of the "family meeting" at which the principal informs all the staff at the same time.

Parents

They too have a personal interest in the results, and most of them will get part of the story from local media. Many will have a limited knowledge of the mechanics of the testing programs and the instructional implications of the scores—and perhaps even a limited desire to learn all the details about them. Nevertheless, parents need to be informed directly. The message should be concise and in language they can understand. It should leave no major question open as far as their child's or their school's performance. If the scores are low, the parents will be concerned, so the report should refer to plans already under way to correct deficiencies. The means of delivering the message depends on the size of the district. Ideally, it should be done at an open meeting or forum where staff members can answer questions on the spot. If that's not possible, school or district media—newsletters, special reports sent home with students or mailed if feasible, will work. The district should never count on the local newspapers to tell the full story the way it wants it told.

Nonparent taxpayers

These citizens are not as personally affected by the results as the parents, although they are interested in the quality of their local schools. They will probably get all the information they need from the local media, but the school system should protect itself from misinterpretations by reporting the results in a community newsletter. The information should be concise and clear, and communicated to nonparent taxpayers as soon as possible after the general release.

The press

Newspaper reporters want all the information they can get, and will often agree to delay publication while they get the chance to absorb it. (See subsequent section on news media). What is important here is that they get easy-to-understand explanations along with the scores, and that some knowledgeable person on the staff is detailed to answer questions without delay. The well-developed news release assumes the media have no background in the subject at all, although individual reporters may have been in the schools often. Important: deadlines are critical, especially where there is competition. If a weekly newspaper's deadline is Wednesday afternoon and the material is released Thursday morning, the resentment could color the coverage.

Television and radio reporters

Much of what was said about the press goes for these reporters, with a few additional considerations. Television reporters want pictures; radio reporters want sound. If the test reports can be made visually attractive, that will help; otherwise, the television station can create the graphics if given enough time. The superintendent should be prepared to be interviewed on the air, because these reporters would not be doing their jobs otherwise. Thus the timing could be critical and must be considered in the plan.

The analysis of the situation should have shown that the news media do not really constitute a public as much as a means of reaching other publics. They presumably have no axe to grind, so the message should be directed to the community at large, not to the reporters and/or editors.

And so forth . . .

Again, it would be impossible to list here the unique characteristics of all publics in a given community and what the considerations should be for each. The analysis phase can take place many months before the event itself, giving the school administration enough time to consider all the implications and plan accordingly.

What, for example, is the relationship with the local government or the legislature? Do these groups expect to put the school system on the defensive? If so, the message and the means of delivering it should anticipate it. How will the unions react, whether the scores are low or high? (Has the term "merit pay" been heard recently in the district?)

It should be noted that where budget and staff permit, many districts conduct public opinion surveys to determine the attitudes and informational needs of various segments of the community. If that is possible, even on an informal unscientific basis, it would be far better than merely guessing what the reaction will be and what are the concerns among staff and community.

In the final analysis, it is likely that many of these audiences could be listed under "nonparent taxpayers" for the sake of the announcement, but it is always advisable to think carefully about each one first before moving on to the next step: planning the communication itself.

Step 2:

Planning

After the analysis is complete, it is time to design the communications (note the plural). First, however, it is essential to think about what the school district wants to accomplish with the results. Too many people merely send out messages without considering that every communication has the potential of provoking an attitude or behavior change. The planning phase is to ensure that the change is a desirable or intended one.

While obviously the goal of the *assessment* program itself is to improve student achievement, the specific goals of the *communication* program may be the following:

- To raise awareness.
- To gain compliance.
- To elicit support.
- To encourage use.

The goals for teachers and parents are obviously different and the communications should be tailored accordingly. For example, making seasoned teachers "aware" of the testing program may be an insult. They have been involved in it for years. New teachers, however, may need a good orientation. The superintendent will earn their "compliance" and the "support" of the principals by detailing how they might be "using" the information in their classrooms in the future.

Another important purpose of the planning phase is to coordinate the activities of the staff members who might be involved in communicating with various publics and to define the terms that will be used. Everybody should be aware that some of the terms used by assessment professionals mean something different to laypeople, so this is the time to establish what terms will be employed to describe these reports. Whatever conclusions are to be drawn, everyone who speaks for the school district should have basically the same message.

This phase of the communications process should never be sold short. It is a chance to anticipate various reactions to the information being released and to plan accordingly. If a public outcry or press criticism puts the school system on the defensive, it may have a hard time recovering.

Public relations professionals say that a communication will be only as effective as the planning that preceded it. Or put another way: Those who know which way the wind is blowing don't whistle into it. Yet it's amazing how many school people try it and get blown away.

Step 3:

Communicating

The time has finally arrived to start communicating. Based on the foregoing analysis and the unique characteristics of the testing program, the challenges should now be clear. In the unusual district where the scores are going up and the taxes are going down, a straightforward and positive message will be easy to write. But few administrators will be that fortunate. In fact, at the local school or district level, it can be a no-win situation no matter how the scores turn out. If they're low, the community reaction is predictable: "What's wrong?" If they're high, somebody is bound to ask, "How come they weren't higher?" This is a reminder of how risky it can be to release test scores without preparing the community over the long haul.

The analysis of the publics and what the scores mean to each of them should be recalled here. Some, like the parents and the teachers, take them to heart, but they are important to everyone in the community because they are one measure of the schools' effectiveness. It is never a good idea to make excuses or try to downplay the relevance of the testing program. It should be faced squarely and related to the bottom line: The education of children.

Every plan is unique

A really good communication plan will work for one situation and one alone. Nobody else will have the identical combination of scores and trends, as well as the same levels of satisfaction or skepticism in the community and cynicism or support among the staff.

The suggested communication steps that follow point up some potential pitfalls and timely tips for avoiding

them. They can be plugged into the planning process and adapted to specific situations. Several of them, in fact, were described at the beginning of this chapter. Many of them, as it turns out, were steps in the communication process that took place many months before the actual "event" itself—steps that make the immediate task immeasurably easier.

From the horse's mouth

Another phenomenon verified by the annual Gallup Poll on Public Attitudes Toward Education is that people who get most of their information directly from the schools themselves are more supportive than those who rely on the media. In the case of assessment results, this suggests that schools should concentrate on getting information directly to the community, and not depend solely on the media to do it for them.

A school district that maintains open lines of communication throughout the year with newsletters, open houses, volunteer programs, partnerships, and so forth, can turn the annual "event" of releasing test scores into a routine occurrence. In fact, virtually any public relations crisis can be mitigated by a continuing program of information and involvement—not just with parents but with other segments of the community as well.

It is too late to prepare the community for a "big announcement" a week before the headlines appear, but unfortunately that's what too often happens with state-run accountability programs. The state department of education holds a news conference and releases columns of numbers. The newspapers print all the numbers and everybody focuses on how the districts rank from top to bottom. Those on the top can bask in the glory, but because of the shock effect, those at the bottom have a hard time expressing pride that their scores are up from the previous year. While the state agency may have succeeded in getting publicity, it may have actually hampered the local school districts from making maximum use of the results to improve instruction.

Being first when you're last

Such a "who's on first" scenario took place in Maryland in 1990 when the newly created state School Performance Program scheduled the release of its first annual "report card" of school districts. Despite the protestations of many school systems who felt they could soften the impact by releasing their own data first, the State Department of Education insisted on enforcing its embargo to get maximum press attention with its district-by-district report.

A week before the scheduled news conference at the state offices, the Baltimore Sun appeared with a story about the performance of the Baltimore City schools in the state program. A few days earlier, forums had been held all over the city where school staffs explained to parents and other citizens what the report meant and what the school system proposed to do about its shortcomings.

The city's school administrators had known that Baltimore would come up last in the state and felt that its citizens deserved an explanation. They denied they had deliberately broken an embargo. "We had an ongoing school-by-school accountability program and we decided a year ago to use the state model locally," explained Deputy Superintendent Edward Andrews. "What we released was our own report. It just happened to contain most of the same data."

That did not change the fact that Baltimore finished statistically at the bottom, but it allowed school officials to soften the impact on its citizens by explaining what the data really meant, how the school system was already dealing with the problems, and why a statistical ranking was inappropriate for that type of report. Andrews is convinced that much of that information would have been overshadowed if he had waited for the much ballyhooed statewide release.

It's crunch time

The Baltimore incident is one example of a situation in which a school district participating in a standardized assessment program, despite an ongoing public relations effort, must still deal with a highly publicized news "event." Even though the community may be familiar with the assessment program, in many cases the state collects and compiles the data, leaving the local district little time to study the results and plan its communication strategy.

Even where local superintendents are less than enamored of state standardized testing programs—and many respondents to the AASA survey hinted that may be the case—a state announcement cannot fail to have an impact on the community. So even if the groundwork has not been laid as recommended earlier in this chapter, there are still ways for the local district to handle the situation.

Here are some suggestions for dealing with the release of standardized test scores in a hypothetical situation where the school district has not instituted a continuing program of communication and involvement:

Start with the staff

As soon as the data are available, they should be distributed to the school and central office staffs. Even though principals and teachers were involved in the administration of the tests, they will not necessarily understand what the assessment results mean in the context of the state and the region. The purpose of these briefings is to make sure that *all* staff members understand the real picture so that they

How to Report on Performance Assessment

Reporting the results of a performance assessment can be very complex because the scores are an entirely different concept from the traditional percentiles, stanines, or grade level measures. And because of the wide range in knowledge and background among the target publics—principals, teachers, and parents—it would be impossible to make a single report serve all of them well.

The California Assessment Program came up with a solution in its writing assessment program: Three separate reports, all based on the same data but including the type of information needed by each of the publics in each school tested. These reports were issued for the first four years of the program, but felt the pinch of the budget cutback in 1991.

There were three booklets, all with the same general format and cover design except for the top line that indicates whether the report is directed to principals or teachers, or is simply the "School Report" for the district.

The principal's report, assuming a high degree of understanding of the program, is only four pages long. It includes a special boilerplate section on interpreting the report and one on how to enhance the school's writing program, along with phone numbers of people to call at the State Department of Education for further help. The test report itself shows the score point distributions for the school and the district and the weighted averages in each of the eight section of the writing test (autobiographical incident, evaluation, firsthand biography, observational writing, problem solution, report of information, speculation about causes or effects, and story).

The second report is sent to every teacher in the grades tested, and contains 10 pages, most of it explanations of each section of the test—what is expected of the student and in general, and what each score (from 1 to 6) represents.

The school/district report, which is distributed to district and county offices and one for each school, is 24 pages long, full of background information about the testing program and the individual tests, and features a description of what each numerical score represents in each of the eight subtests, plus scaled scores to permit comparisons from year to year, among content areas and across grade levels.

ACHIEVEMENT IN AUTOBIOGRAPHICAL INCIDENT

Description of the Rhetorical Effectiveness Score	SCORE POINT	YEAR	SCHOOL		DISTRICT		STATE	
			%	Cumulative %	%	Cumulative %	%	Cumulative %
The student								
* produces a coherent, dramatically realized narrative that relies on a variety of appropriate strategies * successfully orients readers by presenting context, scene, and people * states or implies significance of the incident	Exceptional Achievement 6	86-87 87-88 88-89	3 4 2	3 4 2	3 4 3	3 4 3	3 4 4	3 4 4
* produces a coherent, engaging narrative that successfully orients readers * lacks insight and range of strategies of the writer of a 6 essay * states or implies significance of the incident	Commendable Achievement 5	86-87 87-88 88-89	10 11 8	13 15 10	10 11 10	13 15 13	10 11 11	13 15 15
* produces a well-told incident but lacks the momentum and interest of a 5 or 6 essay * orients readers adequately * states or implies significance but usually adds it to the end of the narrative	Adequate Achievement 4	86-87 87-88 88-89	31 32 29	44 47 38	30 32 29	43 47 42	29 30 29	42 45 44
* produces a narrative that is either brief or rambling * includes only a limited statement or implication of significance	Some Evidence of Achievement 3	86-87 87-88 88-89	37 36 39	81 83 75	38 36 36	81 83 78	36 35 35	76 80 79
* produces a general or fragmentary narrative * includes little if any orientation or reflection on significance	Limited Evidence of Achievement 2	86-87 87-88 88-89	14 13 18	95 96 93	15 13 15	96 96 93	18 16 16	94 95 95
* responds to prompt but with only the hint of a narrative	Minimal Evidence of Achievement 1	86-87 87-88 88-89	5 4 6	100 100 99	5 4 5	100 100 98	5 5 6	100 100 100

BEST COPY AVAILABLE

can respond appropriately to parents or press. And it would not hurt to remind them of the instructional implications of the assessments and how the district plans to use them to bolster student achievement.

Then tell the parents

Parent forums should have been scheduled at each school to coincide closely with the release date. The principal or an authoritative central office staff member should explain the program and the results, distributing illustrative materials and—where appropriate and possible—the children's test scores. In the latter case, time should be available for a concerned parent to talk with a teacher about a child's score and what it really means.

The important thing to remember in this kind of meeting is that the parents' scope of interest may vary. Ron Volpatti, director of pupil services for the Brownsburg, Indiana, Community School Corporation, learned this somewhat to his surprise when he scheduled a "Parent Awareness Night" to answer questions about the statewide testing program.

"We had 90 percent of the parents show up to pick up their kids' scores," he reported, "and then all but a handful left once they saw their kids had passed." The subsequent program he moderated played to a near-empty house. Yet, it was a success. The parents with questions had a chance to ask them. The others were satisfied because Volpatti had made sure they received information about the tests throughout the year through newsletters and teacher conferences.

That was especially important because it's not easy to make testing concepts clear to the average citizen, and it's even harder to do it in the context of the actual test report because people will focus on results rather than explanations.

What Brownsburg does is to anticipate the actual reporting of the standardized test results by publishing background material and tips in the community newsletter. That way parents are prepared for the report when their minds are not clouded by worry or controversy.

What Volpatti also learned is that it doesn't pay to overload parents with a lot of peripheral information or debates over norm-referenced versus criterion-referenced measures. "They mostly care about how their kids did and what it means to their school career."

Put it in writing

Something as potentially complex as a standardized assessment program can be very difficult for a lay person to absorb from a verbal presentation. The gist of the information should be compiled into a written form, along with explanatory visuals, to help people understand it.

The greatest impact would be a summary in the school newsletter sent home to the parents, highlighting that school's results, comparing it with the rest of the district, and showing how it fits in with the rest of the assessment information the parents have been hearing about throughout the school year. In addition, the school system should report the districtwide picture in its own community newsletter or, as an alternative, a special Test Report to the Community.

In either case, while it is important to include the numbers, it is also important to put the program into perspective and clarify some of the terms that people might not understand.

If the superintendent is fortunate enough to have access to space in the local newspaper for his or her own column, this provides still another opportunity to tell the story straight.

In all of these efforts, by taking the initiative and describing the program in its own words, the district is reducing the risk of misinterpretation by the media and/or misunderstanding by parents and citizens.

The basic message

As in all communications during this period, it is advisable not to stray too far afield from the basic message, which is:

- These are the test results.
- Here is where we did well.
- Here is where we didn't do as well.
- This is what we're doing about it.

Even a cursory analysis of the various publics will indicate the level of knowledge, needs, and understanding of each, so that this basic message can be fleshed out where appropriate. What is important, however, is that this is not the time to debate the wisdom of the state testing mandate or the validity of the scores. That will sound like a cop-out. Better yet, relate this information to the district's ongoing assessment program and show how the schools are in control of the situation.

The basic language

Again as in all communications during this period, it is essential to make the report understandable to the audience. It would be easy to take the test report as compiled for the superintendent or school board and retype it for the community newsletter. Easy, but risky.

Educators who have been steeped for years in the jargon of their trade often don't realize that their words often either mean nothing—or something entirely different—to lay people. A test report can become very troubling to the average citizen if not understood, so it is up to the school district to make it clear without talking down to the reader.

And now, the media . . .

While we have left the discussion of the media for last here, the actual release should not be delayed any longer than absolutely necessary. It may be that the state authorities have already released their information, so the local media just want the local angle—and won't wait for it.

The challenge of dealing with the media varies dramatically from district to district. Some don't even have local newspapers, radio or television, while others have all three. As for the local media, some are highly supportive while others are highly suspicious. Some may take the education beat seriously and be well-informed of this and other school programs. And some may be judgmental without the facts, and others downright nasty. But every one of these communities should remember this old adage of school public relations professionals: The news media will always get the last word.

Without getting into the justice of the situation, suffice it to say that it does not pay for a school or school district to get into an argument with the media. This is especially true of a local newspaper, which can make hay over such a feud for many years.

Actually, it works to the advantage of the schools to be cooperative and open with the media. Openness plus honesty equal trust. In the case of releasing assessment results, it is risky just to send out a packet of statistics and hope for the best. Where possible, some districts invest time in meeting individually with reporters covering the story. It would be beneficial all the way around if district staff could actually serve as consultants to the media, helping reporters better understand the stories they are covering.

In the case of assessment results, a news conference allowing plenty of time for the reporters to absorb the information and ask questions is more apt to bring positive coverage. In fact, this conference can be scheduled about the same time as the staff briefing. It might even be possible to put a two-or three-day embargo on the release so that the board and staff don't hear the news first on television. As long as it's no longer than that, the papers usually will not only honor the embargo but will appreciate it. It gives them the chance to digest the material and get back to with questions. (Of course, someone should be available to answer those questions.)

The basic rules of media relations

Even where the scores are low, a school district can survive such a news conference, and subsequent dealings with the media, by observing these basic rules:

- Give the media everything that is available. Do not hide anything. They will probably find out anyway, and then the trouble will be magnified.
- Never belittle or downgrade the program or any of the data. However, it is all right to point out the limitations as well as the benefits of the testing program, whether the scores are high or low.
- Highlight the most important points and explain their implications in the ongoing effort.
- Prepare charts or tables to hand out. This makes it possible to arrange the statistics in a more easily understood format without distorting the data.
- Avoid getting on the defensive.

While it's impossible to write an all-purpose script to fit every situation, consider the following scenario for a news conference adhering to the above rules. A brief summary has already been prepared and distributed to the reporters along with the scores, but rather than read from the summary, the superintendent reviews it orally:

"On the whole, we are very pleased with the results of the assessment program. The report does, however, point up some areas where improvement is called for, such as in fifth-grade reading and seventh-grade math. Believe me, ladies and gentlemen, as soon as we received these results we went right to work in these areas. We've got specialists working to get to the roots of these problems and solve them."

By bringing up the problem areas first, the superintendent has disarmed the situation and has set the stage for a positive approach. Even if the results are not especially gratifying, there are ways to present the information in a positive light. After all, whatever the results are, they will be used to make needed improvements in the instructional program. So even if the news is not particularly good, it is welcome. The school district, it turns out, is taking all this in stride, has the situation well in hand and is pleased to share it with the world.

Front and center

Incidentally, it was no accident that the superintendent took charge of the news conference. The superintendent can and should defer to the experts for technical questions, but the superintendent is the spokesperson the reporters want to hear from on test results. After the reporters get the initial quotes from the superintendent, they will be happy to deal with other administrators, like the testing coordinator, who will be able to answer with more detail.

Lights, camera . . .

A few words about television and radio reporters. It is the nature of their job to simplify. A superintendent who protests, "It's impossible to boil all this down to 30 seconds" will find to his or her dismay that it is not impossible

at all. The only question is who is going to do the boiling—the superintendent or the reporter?

Media-wise superintendents know how to make their points in 20-to 30-second "sound bites." They don't drone on for minutes at a time, challenging the video editors to find a spot—any spot—where they can chop it off gracefully. In television, editing often comes down to seconds, so it pays to be prepared before agreeing to an interview.

It is often helpful for the communications director or other staff member to ask some hard questions to "rehearse" those administrators who will be interviewed by the media. Answers can be critiqued for clarity and accuracy before they actually go before the cameras or microphones.

In those communities with television or radio stations, it can be very valuable for the superintendent to appear on local interview programs and news or talk shows. This personal approach can give the message more credibility with the audience.

Step 4:

Evaluating

The final briefing is over and the last reporter has left the room. The job is over. Right? Wrong!

Communication is analogous to assessment. Just as scoring the tests is not the end of the assessment program, reporting the scores to the public is not the end of the communications task. There is one last step in the public relations process that will make all the difference when the time comes once again to report the results of an assessment program. That step is evaluating, and it should be as structured as the first three.

It will not be too difficult to determine how the school district fared with the media, but even here the size of the headlines isn't the only factor. Did they understand the issues involved? Did they get the implications straight? If they did not, what can school officials do to make sure they get things right the next time?

It will do no good to embark on a nit-picking war over interpretations. Instead of quibbling over concepts, it would be better for the school administrator to reach out and offer to help a reporter observe first-hand what all this means in the classroom. This can be the start of the recommended year-round relationship in which the reporter can learn what assessment is all about and how standardized tests are only a small part of the process.

For the most part, if there are serious parent concerns, the schools themselves will feel the impact directly. The central office will hear it from the rest of the community, including even government and business officials. All of the repercussions should be recorded as they happen to determine whether the reaction was due to a lack of understanding, a deficiency in the process, or to the way the situation was handled.

When the phone rings, somebody should keep a log of related calls with notes about the questions asked. Letters should be filed and the schools canvassed for reports of the reactions there. Clippings of the newspaper stories and tapes of radio and TV reports should be kept for future reference. Reporters might be asked for their advice on how the district could be even more helpful.

If there were big holes in the communications effort, they will surface again and again in the form of the same misunderstandings. This will provide strong clues for the next analysis phase.

The analogy

There is a strong analogy between a communication program and classroom instruction. The evaluation phase is to the communication program what assessment is to instruction. Both communication and instruction are ongoing. Evaluation and assessment are periodic, and they are used to modify and refine communication and instruction.

So while a communication program may have a beginning or an end—like a school year—it really doesn't stop. And while it is important to sit down specifically to analyze the results of the program, it could well be that the first steps in the plan have already taken place.

To repeat the concept that led off this section: Communication, like assessment and school improvement, is a never-ending process.

Preparing the Parents

Following is a sample article in the form of a news release that can be adapted as necessary for a specific situation. It is written in everyday language to give parents a better idea what to expect when their children's standardized test scores are released. The article was produced by the Illinois Association of School Boards (IASB) for subscribers to its School PR Service, with thanks to the National PTA and the Educational Testing Service for the information in their flyer, "Plain Talk About Testing." It is reproduced here with permission from IASB.

Your Child's Test Scores: What Do They Mean?

Schools use all kinds of tests to measure your child's progress and potential. Some are easy to understand—such as the classroom tests that measure how well your child has learned mathematics, science, and other subjects.

But it can be difficult to understand the meaning of the scores on other tests. Some of the tests schools use include:

- Standardized tests, such as the ***(fill in standardized tests)*** used in ***(fill in name of your school or district)***. Standardized tests allow comparisons of students in different school districts around the country.
- Achievement tests measure how much a student has learned in a subject. Classroom tests, including final examinations, usually are achievement tests.

Some achievement tests compare what a student has learned with other students around the country; others measure what a student knows without making comparisons.

- Aptitude tests measure skills and abilities that are useful in learning specific subjects. Because they show which skills students excel in, they can be used to predict how well students will do in school, to help in career planning, or to identify gifted children.

However, if a student does not do well in certain skills tested, that does not mean he or she won't do well in school. It just means that certain skills may be harder for some students than for others.

College admissions tests usually include portions to test both aptitude and achievement, because colleges want to know both what a student has achieved and what that student is capable of achieving.

Tests are scored in different ways, too:

- Percentiles compare a student's scores with scores of a norm group. The percentile shows what percent of those in the norm group the student scored above. For example, a percentile of 75 means the students did better than 75 percent of those in the norm group. It does not mean that the student got 75 percent of the questions right.
- Grade equivalent scores are reported as two numbers separated by a decimal point. The first number stands for a grade in school and the second for a tenth of a year. It shows a performance level that is typical for students in that grade at that time of year. A grade equivalent score of 5.5, for example, means the student scored about as well on that test as the average student who is halfway through fifth grade.

It is important to remember that a grade equivalent score is a score that is in the middle of the class. Therefore, half of the students in a school are likely to fall below that score.

- Stanine scores are single-digit numbers. The lowest is one and the highest is nine. You can think of scores of 1, 2, and 3 as being below average; 4, 5, and 6 as average; and 7, 8, and 9 as above average.

The most important thing to remember about tests is that no single test is all-important. If a student does badly on one test, but his or her other test scores are good, there are several possible reasons. One reason is that some tests are more reliable than others, and no test is entirely reliable.

How well a student will do on any given test also depends on whether the test happens to hit on a lot of questions the student could answer; how well a student feels that day; and how lucky the student is at guessing the answers he or she doesn't know.

That doesn't mean tests are worthless. They provide valuable information for students and schools. It just means that you shouldn't be too concerned about any single test score. Look instead at the pattern or a student's test scores and his or her accomplishments. That pattern will tell you how well your child is doing far better than any single test score ever could.

To understand your child's test scores, be sure you:

- *Understand what the test measures.*
- *Understand what the test means.*
- *Consider how well the child is doing in general.*

If you are concerned about a test score, contact his or her teacher or counselor. They'll be glad to answer your questions.

(Copyright 1986 by Illinois Association of School Boards. Reprinted with permission.)

CHAPTER TEN

More Than Just Test Scores

One of the outcomes of the reevaluation of educational assessment is the realization that there is a lot more to school performance than test scores. Responding to demands for "accountability," several states have broadened the definition of school quality.

As we have seen, when the federal government began to finance a wide variety of offerings in America's schools about 25 years ago, the *quid pro quo* was that the recipients prove the money was being put to good use. The term "accountability" became part of the vocabulary of virtually every school system, and in most cases it was tied directly to scores on standardized norm-referenced achievement tests.

In the mid-1980s, when reform efforts were launched around the country in response to critical reports on the state of American education, once again norm-referenced standardized tests were called upon to measure their success. But along with the rapid increase in such testing came a rising chorus of opposition.

As the decade came to a close, many standardized tests became criterion-referenced and even performance-based, but the critics persisted. Standardized testing may have its place, they said, but it fails in two very important areas:

- It does not relate directly enough to classroom practices to guarantee improved teaching and greater learning.
- It can be very misleading when used as the definitive measure of a school's performance.

Much of this book has been devoted to new forms of assessment that can be used to improve student achievement while at the same time producing accurate reports on school performance. The question still remains, however: Are test scores enough to prove that schools are doing their job?

While the topic of "accountability" is a subject worthy of its own book, there are many promising movements under way around the country that deserve mention here. Here are some examples of how reports to the public are being expanded to demonstrate school performance with more than just test scores.

A new kind of report card

Like many other states, Maryland launched a full-scale "accountability" effort in the early 1970's hitched entirely to norm-referenced standardized testing. Using first the ITBS and then the CAT, every school system tested every student in grades 3, 5 and 8 every year. The percentile scores, school by school, were dutifully reported to the public and delivered, as mandated by law, to the state legislature.

At the same time, in a separate program, statewide minimum competency examinations labeled "Functional Tests" were administered to all 9th graders in reading, writing, mathematics and citizenship. Students who failed them could not graduate, but they could retake them periodically until they passed. Debate continually surrounded these tests as to whether they were (1) meaningful and (2) fair.

As reported earlier, when Governor William Donald Schaefer took office in 1987, he noted that the scores of Maryland students on the standardized tests were well above national norms in reading and math, as they were on the SAT's. But did those rosy numbers translate into excellence in the state's public schools? Governor Schaefer thought not. So in 1987 he established a high-powered Governor's Commission on School Performance to answer questions like these:

- How can a state have high test scores and high absentee and dropout rates at the same time?
- If scores are so high, why do employers complain that many high school graduates have marginal skills at best?
- How can the true quality of each school be measured, and how can the state ensure that school performance attains—and maintains—that quality?

One of the first things the commission did was to dismiss the existing testing program as ineffectual in keeping track of student performance. In its final report two years later, it recommended sweeping changes in the state's accountability program.

The "Vital Core"

The commission was insistent that there were many measures of a school's performance in addition to those that can be measured on a test. It spent many months debating what it called a "vital core of student achievements in which statewide measurement is possible and desirable." This was the information to be compiled for each school in the state:

- ***Post-secondary attainment***—The accomplishments of all Maryland high school graduates, those who go on to higher education as well as those who go directly into the work force.
- ***Academic attainment***—The percentage of students completing a rigorous academic program (based upon what the state university system will count toward its course entry requirements).
- ***Minimum progress toward graduation***—The percentage of students who pass competency tests required for graduation by the end of the tenth grade.
- ***Attendance and dropouts***—The percentage of average daily attendance and the percentage of students dropping out of the school.
- ***Student knowledge***—Tested level of performance in reading, mathematics, writing, science and social studies. The recommended testing program was discussed elsewhere in the report, which specifically condemned norm-referenced tests. Here the report called for a method that
  - Measures fundamental, important things in the curriculum.
  - Measures more than a narrow, low-level set of objectives.
  - Does not encourage rote teaching of material covered by the test.
  - Holds schools, in addition to students, accountable for achievement.

The commission did not recommend a specific testing program, other than to urge that it be "very different" from the existing accountability program using the California Achievement Tests. Its final report said this:

"First we would require reports on each school, not just on those schools with students tested at specific grade levels. Second, the school report would give the percentage of students actually meeting each standard, rather than the group's average score. A child's poor performance—and need for help—should never again be hidden behind a school's high average scores, and a high performing student should not be denied recognition by being obscured among low average scores. Third, we would require school systems and schools to examine factors that may influence student achievements."

Other factors

The commission was seeking to isolate those factors that represent both conditions under which the educational process takes place and efforts made to support and manage the educational process. Many factors were debated over the months, but these were the ones selected "that are often thought to influence student achievements or, at least, to help describe efforts made or the magnitude of the tasks schools face:"

- ***Disadvantaged:*** The percentage of students who receive free or reduced-price lunches.
- ***Mobility:*** The percentage of students enrolled who enter and leave a school each year.
- ***Wealth per pupil:*** A district's total taxable wealth per pupil.
- ***Staffing:*** The ratio of student enrollment to full-time equivalent instructional and support staff, expressed separately.
- ***Staff attendance:*** The percentage of average daily attendance of school staff.
- ***Instructional time:*** The total length over a school year of the time allocated to instruction (in hours).
- ***Per pupil costs:*** (As defined consistently in each district by the State Department of Education.)

This list of factors was the source of much debate during the committee deliberations. Mixing inputs with outputs is not common practice. And some members feared that school officials would hide behind them. But the commission retained them with the following explanation:

"We suggest the use of these factors to examine comparable schools as one way to identify exemplary performance—even under the most difficult circumstances—and to identify clues to achieving excellence in other schools in similar circumstances. Information gained from this analysis should be used to inform policy makers, but never be an excuse for poor performance."

Rapid implementation

Governor Schaefer bought the recommendations with a great deal of fanfare, and within a few months the State Board of Education had launched the Maryland School Performance Program (MSPP) and set the state's 24 school districts to work on a feverish pace. The scheduled administration of CATs was scrapped while a new testing program was developed, based upon the concept of performance assessments.

Even before that task was finished, the first school report cards were issued in 1990 as a baseline, showing how each

school system fared against a standard established (admittedly arbitrarily) by the State Board of Education, and against the state average.

Here is a sample of that report, showing a county school system. Note how the other factors are covered on the page headed "Supporting Information."

STUDENT PERFORMANCE

*** BASELINE DATA—SCHOOL YEAR 1989-90**

Assessed Knowledge

Md. Functional Tests (First Administration)	Standard %		Number Taking	Percent Passing	Excellent	Satisfactory	Standard Not Yet Met
	Excellent	Satisfactory					
Reading (Grade 9)	97	95	2,022	96.6		✓	
Mathematics (Grade 9)	90	80	2,046	77.5			✓
Writing (Grade 9)	96	90	2,036	94.5		✓	
Citizenship (Grade 9/10)	92	85	2,048	81.2			✓

Student Participation

Attendance (Yearly)	Standard %		Percent	Excellent	Satisfactory	Standard Not Yet Met
	Excellent	Satisfactory				
Grades 1-6	96	94	95.2		✓	
Grades 7-12	96	94	93.2			✓

Dropout Rate (Yearly)	Standard %		Percent	Excellent	Satisfactory	Standard Not Yet Met
	Excellent	Satisfactory				
Grades 9-12	1.25	3.0	4.0			✓

Student Attainment

Promotion Rate (Yearly)	Standard %		Percent	Excellent	Satisfactory	Standard Not Yet Met
	Excellent	Satisfactory				
Grades 1-6	98	96	98.5	✓		

School Improvement Highlights

The student scores on the functional tests continue to show steady, significant improvement. The results in each of the four areas tested should meet either the excellent or satisfactory standard during the 1990-91 school year.

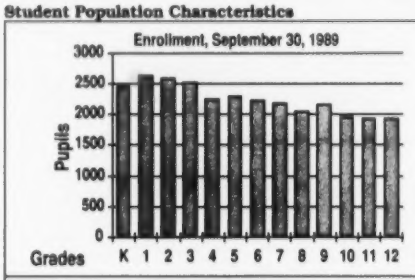
High percentages of promotion, very good attendance, and a relatively low dropout rate attest further to the effectiveness of the Harford County Public School System.

*In future years, this baseline data will be used to chart progress toward and beyond the standards. Additional areas will be added in 1991 and 1992.

SUPPORTING INFORMATION

Harford

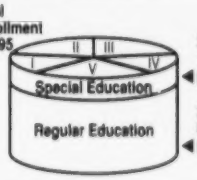
Student Population Characteristics



Mobility

Students	Number	Percent
Entrants	2,607	8.6
Withdrawals	3,089	10.2

Students Receiving Special Programs



3,458 pupils — 11.8%

25,837 pupils — 88.2%

Special Education

Level	Number	Percent
I	381	11.0
II	1,459	42.2
III	794	23.0
IV	583	16.8
V	241	7.0

	Number	Percent
Non English/Limited English Proficient	130	.44
Chapter 1	2,516	8.6
Free/Reduced Priced Meals	3,592	12.3

Other Factors

Wealth per Pupil	$120,968
Per Pupil Expenditure	$4,191
Instructional Staff per 1000 Pupils	59.4
Professional Support Staff per 1000 Pupils	8.4
Instructional Aides per 1000 Pupils	8.4
Average Length of School Day for Pupils	6:30
Length of School Year for Pupils	180

Notes

While the levels of pupil achievement are among the highest in the State, per pupil expenditure and staffing ratios are among the lowest when compared to other public school systems.

BEST COPY AVAILABLE

In subsequent years, the results of the testing program in the academic subjects will be included, and the reports will be issued on a school-by-school as well as county-by-county basis.

School profiles in Delaware

The impetus in Delaware also came from the Governor's Office. Governor Michael N. Castle felt that the public needed more information so they would understand what the state's schools were all about.

State Superintendent William B. Keene obliged, and despite some trepidation among school administrators around the state, the first individual school profiles were distributed in 1990 for all of Delaware's 25 high schools and three vocational schools. The reports, compiled by the State Department of Public Instruction, took the form of a small 16-page booklet that painted a picture—some statistical and some narrative—of the school district, the school, and the students.

For example, it opens with the district's demographics and financial indicators, then traces school enrollment trends, reports on a survey of future plans of current seniors, and shows the dropout rate.

The section on student outcomes describes the results on the Stanford Achievement Tests in grade 11 as well as on the Scholastic Aptitude Test. In both cases, the school's scores are compared with the state average.

After a two-page list of all the school's course offerings, the book concludes with two pages prepared by the school itself to describe the accomplishments of the staff and students during the previous school year and the discuss its commitments for the future.

These reports were to be extended the following year to cover all middle schools, and eventually to include every school in the state.

The fears of some of the local school officials turned out not to be justified, recalled Ambrose W. Hagarty, the department's public information officer. They had worried, he said, that publishing certain test score information—especially the Scholastic Aptitude Tests in a state with high private school enrollment—would cause a public outcry. As it turned out, it didn't.

"You can't assess anything of consequence unless you see how each intersects and is affected by others."

—Theodore Sizer.

A survey of educators and parents, Hagarty said, showed that 90 percent considered the profiles a positive step and found them helpful.

. . . and in Arizona

It was the legislature that forced the issue in Arizona. Its act called "Goals for Educational Excellence" promises to change much of the way the schools in Arizona do business. The law, which required schools to set high goals and measure how well they were approaching them, will force school districts to move away from sole reliance on norm-referenced standardized tests, but, as in Maryland and Delaware, will also introduce into the mix factors other than academic achievement.

When the first School Profiles are published in 1992, they will include those factors that "affect" and those that "reflect" student achievement without testing —things like parent involvement, enrollment in advanced placement courses, etc.

Lois Easton, director of curriculum and assessment, said those factors will be as important as the new performance assessments being developed, because the profile will ultimately become a "school-site management tool." Before it's all over, Easton mused, what looks like merely a new assessment program may turn out to be an "overhaul of the whole system."

. . . and Michigan too

The Michigan legislature used a carrot and a stick in its requirement for individual school annual reports starting in 1990-91. The law, Public Act 25, was very specific how they would include—but by no means be restricted to—standardized test scores. The carrot was a bonus of $25 per pupil for a district that satisfies the requirements; the stick was a forfeiture of 5 percent of all state funds for those that don't.

Briefly, the annual report shall include, but not be limited to, the following components:

1. School improvement plan status report.
2. Student achievement based upon local, state and/or national tests and other measures. This is to include evidence of student achievement as defined by building staff through the school improvement process; other evidence of student achievement is encouraged.
3. Retention and dropout rate report.
4. Specialized schools and placement process.
5. Parent participation percentage in teacher conferences.
6. Accreditation status report.
7. Core curriculum and implementation plan.

To some school administrators, this was the chance of a lifetime to show what they are all about. Cynthia E. Banach, special projects administrator of the St. Clair Intermediate School District, urged constituent school districts and their building level administrators to take this opportunity to create a report that can have far-reaching effects.

On the form she created to help each school compile its data, Banach advised: "Don't be modest. Now is the time to showcase your students, staff, and the programs you have under way. It is also the time to help bring your school community up-to-date on future programming and building needs."

. . . and the beat goes on

While the impetus behind all of the above alternative accountability reports started in the state capital, many local districts are also keeping track of a variety of factors other than test results. Notations on the AASA survey included the following:

- Dropouts
- Attendance
- Mobility rate
- Extracurricular activities
- Community service performed by students
- Parent attitudes
- Number of graduates going to college.

And even the mass media are getting into the act. A special edition of Money Magazine in 1990 asked the question, "Can Your Kid's School Pass This Quality Test?" The "test" involved asking the principal, students, PTA members and other staff members about 50 questions covering a wide range of objective and subjective observations. It eventually adds up to a mathematical score to show whether a school is "at the head of the class," "needs improvement" or "should be avoided."

The concluding section of the article pointed up some aspects of a high school that could well be a part of future accountability reports. Of course, any of the items on the list is arguable, but without getting into the magazine's statistical game or the validity of its factors, here is the list:

- The average class size.
- The number of students per teacher schoolwide.
- The number of students per guidance counselor.
- The percentage of students who drop out.
- The number of students suspended for disciplinary reasons last year.
- The number of students who were not promoted to the next grade last year.
- The percentage of seniors who graduate.
- The percentage of students who go on to four-year colleges.
- The percentage of the students enrolled who are taking calculus or other advanced math courses.
- The percentage of students who took advanced placement courses last year.
- The number of National Merit Scholarship Qualifying Test semifinalists and finalists.
- The percentage of the student body enrolled in college preparatory courses, rather than general or vocational studies.
- The percentage of teachers who have master's degrees or doctorates.

And perhaps one more item should be added—that of accessibility:

"To complete this exam," the article notes, "you will need a fair bit of cooperation from administrators and faculty. If they are reluctant to answer questions or give you access, you should start looking for another school."

Other AASA Critical Issues Reports

Students At Risk: Problems and Solutions. Helps all concerned with the welfare of children and youth understand the forces that bring them to and over the brink of school failure. Offers scores of strategies and programs for meeting the needs of students who are at risk because of neglect, poverty, abuse, an English deficiency, drug and alcohol abuse, because they are academically over- or under-challenged, and for other reasons. (112pp; 1989; $13.95; Stock #021–00213).

Teaching Thinking and Reasoning Skills: Problems and Solutions. Shows how teaching, thinking and reasoning can be incorporated into the curriculum. Reviews the research and explores programs now at work in schools across the nation. (80pp.; 1987; $13.95; stock #021-00175).

Staff Development: Problems and Solutions. This outstanding publication will help you improve your professional development program for all staff. Exemplary programs are included. (80 pp; 1986; $13.95; Stock #021-00164).

Promoting Health Education in Schools: Problems and Solutions. An informative overview of both the problems and prospects for school health education. (68pp; 1985; $13.95; Stock #021-00152).

Raising Standards in Schools: Problems and Solutions. Includes results of the studies for improving standards and tells how to involve key community groups in achieving quality education. (88pp; 1985; $13.95; Stock #021-00145).

Improving Math & Science Education: Problems and Solutions. Includes examples of how districts are improving math and science education; lists of booklets and other materials developed for grades K-12. (92pp; 1985; $13.95; Stock #021-00131).

High Tech for Schools: Problems and Solutions. Details planning models; examines "computer literacy"; and offers advice on hardware, personnel and staff development, software, and uses of computers in administration and special education. Also discusses cable television, electronic networks, and videodiscs. (96pp; 1984; $12.95; Stock #021-00122).

The Role of the Principal in Effective Schools: Problems and Solutions. Details the principal's role as instructional leader and manager; training, selection, and evaluation of principals; and how the superintendent/board help principals become more effective. (96pp; 1983; $12.95; Stock #021-00839(.

Planning for Tomorrow's Schools: Problems and Solutions. Examines the ingredients of effective planning. Offers planning models and checklists; advice on involving staff and community and setting goals and objectives; information demographics, politics, finance, and attracting good teachers. (80pp; 1983; $12.95; Stock #021-00832).

Building Morale . . . Motivating Staff: Problems and Solutions. Identifies morale builders and morale killers. Includes case studies of districts with excellent morale; tips on inservice, communications, evaluation, and managing conflict (80pp; 1983; $11.95; Stock #021-00834).

Teaching Writing: Problems and Solutions. Describes successful writing programs, how to teach students to use the writing process, tips on inservice for teachers, the principal roles, and how to evaluate grade student writing. (96pp; 1982; $11.95; Stock #021-00901).

School Budgeting: Problems and Solutions. Describes making the budget an integral part of the school program; budgeting methods; planning ahead; and primers on finance campaigns, PR, negotiations, and lobbying. (72pp; 1982; $10.95; Stock #021-00900).

Declining Enrollment-closing Schools: Problems and Solutions. Details lessons learned about declining enrollment. Selected as "must reading" by the National School Boards Association. (88pp; 1981; $10.95; Stock #021-00336).

Collective Bargaining: Problems and Solutions. Selected "must reading" by the National School Boards Association. (95pp; 1981; $10.95; Stock #021-00330).

Student Discipline: Problems and Solutions. Selected "must reading" by the National School Boards Association. (80pp; 1980; $9.95; Stock #021-00334).

How Schools Can Save $$: Problems and Solutions. Includes suggestions and case studies on cutting the costs of staffing, fringe benefits, energy, and insurance. (80pp; 1980; $8.95; Stock #021-00333).

Teacher Competency: Problems and Solutions. Selected "must reading" by the National School Boards Association. (80pp; 1980; $8.95; Stock #021-00332).

Keeping Students in School: Problems and Solutions. Offers help to school districts facing staff dismissal in cases of incompetence, declining enrollment, or budget problems. (80pp; 1978; $8.95; Stock #021-00512).

The Competency Movement: Problems and Solutions. Explains the pros and cons of the minimum competency movement. Includes guidelines (92pp; 1978; $8.95; Stock #021-00510).

Discount and Ordering Information

Order from the American Association of School Administrators, 1801 N. Moore St., Arlington, VA 22209-9988. To order a single copy, please add $3.50 for postage and handling. Orders for less than $20 must be accompanied by payment in full. Discount on any one Critical Issues Report: two to nine copies, 10% discount. Write for prices of 10 or more copies. Please indicate stock numbers on orders.